SIAM J. OPTIM.
Vol. 25, No. 2, pp. 829–855

INCREMENTAL MAJORIZATION-MINIMIZATION OPTIMIZATION WITH APPLICATION TO LARGE-SCALE MACHINE LEARNING*

JULIEN MAIRAL†

Abstract. Majorization-minimization algorithms consist of successively minimizing a sequence of upper bounds of the objective function. These upper bounds are tight at the current estimate, and each iteration monotonically drives the objective function downhill. Such a simple principle is widely applicable and has been very popular in various scientific fields, especially in signal processing and statistics. We propose an incremental majorization-minimization scheme for minimizing a large sum of continuous functions, a problem of utmost importance in machine learning. We present convergence guarantees for nonconvex and convex optimization when the upper bounds approximate the objective up to a smooth error; we call such upper bounds "first-order surrogate functions." More precisely, we study asymptotic stationary point guarantees for nonconvex problems, and for convex ones, we provide convergence rates for the expected objective function value. We apply our scheme to composite optimization and obtain a new incremental proximal gradient algorithm with linear convergence rate for strongly convex functions. Our experiments show that our method is competitive with the state of the art for solving machine learning problems such as logistic regression when the number of training samples is large enough, and we demonstrate its usefulness for sparse estimation with nonconvex penalties.

Key words. nonconvex optimization, convex optimization, majorization-minimization

AMS subject classifications. 90C06, 90C26, 90C25

DOI. 10.1137/140957639

1. Introduction.

The principle of successively minimizing upper bounds of the objective function is often called *majorization-minimization* [35] or *successive upper-bound minimization* [48]. Each upper bound is locally tight at the current estimate, and each minimization step decreases the value of the objective function. Even though this principle does not provide any theoretical guarantee about the quality of the returned solution, it has been very popular and widely used because of its simplicity. Various existing approaches can indeed be interpreted from the majorization-minimization point of view. This is the case of many gradient-based or proximal methods [3, 14, 28, 45, 54], expectation-maximization (EM) algorithms in statistics [20, 42], difference-of-convex (DC) programming [30], boosting [13, 19], some variational Bayes techniques used in machine learning [53], and the mean-shift algorithm for finding modes of a distribution [25]. Majorizing surrogates have also been used successfully in the signal processing literature about sparse estimation [11, 16, 26], linear inverse problems in image processing [1, 23], and matrix factorization [37, 40].

*Received by the editors February 18, 2014; accepted for publication (in revised form) January 27, 2015; published electronically April 14, 2015. This work was partially supported by the Gargantua project (Program Mastodons, CNRS), the Microsoft Research-Inria Joint Centre, and Agence Nationale de la Recherche (MACARON project ANR-14-CE23-0003-01 and LabEx PERSYVAL-Lab ANR-11-LABX-0025). A short version of this work was presented at the International Conference of Machine Learning (ICML) in 2013.
http://www.siam.org/journals/siopt/25-2/95763.html

†Inria, LEAR Team, Laboratoire Jean Kuntzmann, CNRS, Université Grenoble Alpes, 655 avenue de l'Europe, 38330 Montbonnot, France (julien.mairal@inria.fr).

In this paper, we are interested in making the majorization-minimization principle scalable for minimizing a large sum of functions:

$$\min_{\theta \in \Theta} \left[f(\theta) \triangleq \frac{1}{T} \sum_{t=1}^{T} f^t(\theta) \right], \tag{1.1}$$

where the functions $f^t : \mathbb{R}^p \to \mathbb{R}$ are continuous, and Θ is a convex subset of $\mathbb{R}^p$. When f is nonconvex, exactly solving (1.1) is intractable in general, and when f is also nonsmooth, finding a stationary point of (1.1) can be difficult. The problem above when T is large can be motivated by machine learning applications, where θ represents some model parameters and each function f^t measures the adequacy of the parameters θ to an observed data point indexed by t. In this context, minimizing f amounts to finding parameters θ that explain well some observed data. In the last few years, stochastic optimization techniques have become very popular in machine learning for their empirical ability to deal with a large number T of training points [9, 22, 51, 55]. Even though these methods have inherent sublinear convergence rates for convex and strongly convex problems [34, 43], they typically have a cheap computational cost per iteration, enabling them to efficiently find an approximate solution. Recently, incremental algorithms have also been proposed for minimizing finite sums of functions [6, 17, 18, 50, 51]. At the price of a higher memory cost than stochastic algorithms, these incremental methods enjoy faster convergence rates, while also having a cheap per-iteration computational cost.

Our paper follows this approach: in order to exploit the particular structure of problem (1.1), we propose an incremental scheme whose cost per iteration is independent of T, as soon as the upper bounds of the objective are appropriately chosen. We call the resulting scheme "MISO" (minimization by incremental surrogate optimization). We present convergence results when the upper bounds are chosen among the class of "first-order surrogate functions," which approximate the objective function up to a smooth error—that is, differentiable with a Lipschitz continuous gradient. For nonconvex problems, we obtain almost sure convergence and asymptotic stationary point guarantees. In addition, when assuming the surrogates to be strongly convex, we provide convergence rates for the expected value of the objective function. Remarkably, the convergence rate of MISO is linear for minimizing strongly convex composite objective functions, a property shared with two other incremental algorithms for smooth and composite convex optimization: the *stochastic average gradient* method (SAG) of Schmidt, Le Roux, and Bach [50] and the *stochastic dual coordinate ascent* method (SDCA) of Shalev-Schwartz and Zhang [51]. Our scheme MISO is inspired in part by these two works but yields different update rules than SAG or SDCA and is also appropriate for nonconvex optimization problems.

In the experimental section of this paper, we show that MISO can be useful for solving large-scale machine learning problems and that it matches cutting-edge solvers for large-scale logistic regression [3, 50]. Then, we show that our approach provides an effective incremental DC programming algorithm, which we apply to sparse estimation problems with nonconvex penalties [11].

The paper is organized as follows. Section 2 introduces the majorization-minimization principle with first-order surrogate functions. Section 3 is devoted to our incremental scheme MISO. Section 4 presents some numerical experiments, and section 5 concludes the paper. Some basic definitions are given in Appendix A.

2. Majorization-minimization with first-order surrogate functions. In this section, we present the generic majorization-minimization scheme for minimizing a function f without exploiting its structure—that is, without using the fact that f is a

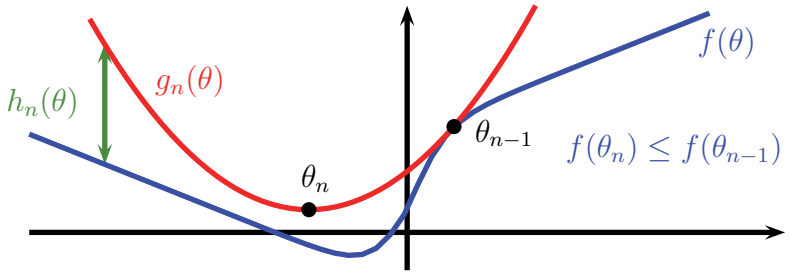

FIG. 1. *Illustration of the basic majorization-minimization principle. We compute a surrogate g_n of f near the current estimate θ_{n-1}. The new estimate θ_n is a minimizer of g_n. The function $h_n = g_n - f$ is the approximation error that is made when replacing f by g_n.*

sum of functions. We describe the procedure in Algorithm 1 and illustrate its principle in Figure 1. At iteration n, the estimate θ_n is obtained by minimizing a surrogate function g_n of f. When g_n uniformly majorizes f and when $g_n(\theta_{n-1}) = f(\theta_{n-1})$, it is clear that the objective function value monotonically decreases.

ALGORITHM 1. BASIC MAJORIZATION-MINIMIZATION SCHEME.

input $\theta_0 \in \Theta$ (initial estimate); N (number of iterations).

1: **for** $n = 1, \ldots, N$ **do**
2: Compute a surrogate function g_n of f near θ_{n-1};
3: Minimize the surrogate and update the solution: $\theta_n \in \arg\min_{\theta \in \Theta} g_n(\theta)$.
4: **end for**

output θ_N (final estimate);

For this approach to be effective, intuition tells us that we need functions g_n that are easy to minimize and that approximate well the objective f. Therefore, we measure the quality of the approximation through the smoothness of the error $h_n \triangleq g_n - f$, which is a key quantity arising in the convergence analysis. Specifically, we require h_n to be L-smooth for some constant $L > 0$ in the following sense.

DEFINITION 2.1 (L-smooth functions). *A function $f : \mathbb{R}^p \to \mathbb{R}$ is called L-smooth when it is differentiable and when its gradient ∇f is L-Lipschitz continuous.*

With this definition in hand, we now introduce the class of "first-order surrogate functions," which will be shown to have good enough properties for analyzing the convergence of Algorithm 1 and other variants.

DEFINITION 2.2 (first-order surrogate functions). *A function $g : \mathbb{R}^p \to \mathbb{R}$ is a first-order surrogate function of f near κ in Θ when*

(i) *$g(\theta') \geq f(\theta')$ for all minimizers θ' of g over Θ, and when the more general condition $g \geq f$ holds, we say that g is a* majorizing *surrogate;*

(ii) *the approximation error $h \triangleq g - f$ is L-smooth, $h(\kappa) = 0$, and $\nabla h(\kappa) = 0$.*

We denote by $\mathcal{S}_L(f, \kappa)$ the set of first-order surrogate functions and by $\mathcal{S}_{L,\rho}(f, \kappa) \subset \mathcal{S}_L(f, \kappa)$ the subset of ρ-strongly convex surrogates.

First-order surrogates are interesting because their approximation error—the difference between the surrogate and the objective—can be easily controlled. This is formally stated in the next lemma, which is a building block of our analysis.

LEMMA 2.3 (basic properties of first-order surrogate functions). *Let g be a surrogate function in $\mathcal{S}_L(f,\kappa)$ for some κ in Θ. Define the approximation error $h \triangleq g - f$, and let θ' be a minimizer of g over Θ. Then, for all θ in Θ,*

- $|h(\theta)| \leq \frac{L}{2}\|\theta - \kappa\|_2^2$;
- $f(\theta') \leq f(\theta) + \frac{L}{2}\|\theta - \kappa\|_2^2$.

Assume that g is ρ-strongly convex, i.e., g is in $\mathcal{S}_{L,\rho}(f,\kappa)$. Then, for all θ in Θ,

- $f(\theta') + \frac{\rho}{2}\|\theta' - \theta\|_2^2 \leq f(\theta) + \frac{L}{2}\|\theta - \kappa\|_2^2$.

Proof. The first inequality is a direct application of a classical result (Lemma 1.2.3 of [44]) on quadratic upper bounds for L-smooth functions, when noticing that $h(\kappa) = 0$ and $\nabla h(\kappa) = 0$. Then, for all θ in Θ, we have $f(\theta') \leq g(\theta') \leq g(\theta) = f(\theta) + h(\theta)$, and we obtain the second inequality from the first one.

When g is ρ-strongly convex, we use the following classical lower bound (see [45]):

$$g(\theta') + \frac{\rho}{2}\|\theta - \theta'\|_2^2 \leq g(\theta).$$

Since $f(\theta') \leq g(\theta')$ by Definition 2.2 and $g(\theta) = f(\theta) + h(\theta)$, the third inequality follows from the first one. ∎

We now proceed with a convergence analysis including four main results regarding Algorithm 1 with first-order surrogate functions g_n. More precisely, we show in section 2.1 that, under simple assumptions, the sequence of iterates asymptotically satisfies a stationary point condition. Then, we present a similar result with relaxed assumptions on the surrogates g_n when f is a composition of two functions, which occur in practical situations as shown in section 2.3. Finally, we present nonasymptotic convergence rates when f is convex in section 2.2. By adapting convergence proofs of proximal gradient methods [45] to our more general setting, we recover classical sublinear rates $O(1/n)$ and linear convergence rates for strongly convex problems.

2.1. Nonconvex convergence analysis. For general nonconvex problems, proving convergence to a global (or local) minimum is impossible in general, and classical analysis studies instead asymptotic stationary point conditions (see, e.g., [5]). To do so, we make the following mild assumption when f is nonconvex:

(A) f is bounded below and for all θ, θ' in Θ, the directional derivative $\nabla f(\theta, \theta' - \theta)$ of f at θ in the direction $\theta' - \theta$ exists.

The definitions of directional derivatives and stationary points are provided in Appendix A. A necessary first-order condition for θ to be a local minimum of f is to have $\nabla f(\theta, \theta' - \theta) \geq 0$ for all θ' in Θ (see, e.g., [8]). In other words, there is no feasible descent direction $\theta' - \theta$ and θ is a stationary point. Thus, we consider the following condition for assessing the quality of a sequence $(\theta_n)_{n\geq 0}$ for nonconvex problems:

DEFINITION 2.4 (asymptotic stationary point). *Under assumption* (A)*, a sequence $(\theta_n)_{n\geq 0}$ satisfies the asymptotic stationary point condition if*

$$\liminf_{n\to+\infty} \inf_{\theta\in\Theta} \frac{\nabla f(\theta_n, \theta - \theta_n)}{\|\theta - \theta_n\|_2} \geq 0. \tag{2.1}$$

Note that if f is differentiable on $\mathbb{R}^p$ and $\Theta = \mathbb{R}^p$, $\nabla f(\theta_n, \theta - \theta_n) = \nabla f(\theta_n)^\top(\theta - \theta_n)$, and the condition (2.1) *implies that the sequence $(\nabla f(\theta_n))_{n\geq 0}$ converges to* 0.

As noted, we recover the classical definition of critical points for the smooth unconstrained case. We now give a first convergence result about Algorithm 1.

PROPOSITION 2.5 (nonconvex analysis for Algorithm 1). *Assume that* (A) *holds and that the surrogates g_n from Algorithm* 1 *are in $\mathcal{S}_L(f, \theta_{n-1})$ and are either*

majorizing f or strongly convex. Then, $(f(\theta_n))_{n\geq 0}$ monotonically decreases, and $(\theta_n)_{n\geq 0}$ satisfies the asymptotic stationary point condition.

Proof. The fact that $(f(\theta_n))_{n\geq 0}$ is nonincreasing and convergent because bounded below is clear: for all $n \geq 1$, $f(\theta_n) \leq g_n(\theta_n) \leq g_n(\theta_{n-1}) = f(\theta_{n-1})$, where the first inequality and the last equality are obtained from Definition 2.2. The second inequality comes from the definition of θ_n.

Let us now denote by $f^\star$ the limit of the sequence $(f(\theta_n))_{n\geq 1}$ and by $h_n \triangleq g_n - f$ the approximation error function at iteration n, which is L-smooth by Definition 2.2 and such that $h_n(\theta_n) \geq 0$. Then, $h_n(\theta_n) = g_n(\theta_n) - f(\theta_n) \leq f(\theta_{n-1}) - f(\theta_n)$, and

$$\sum_{n=1}^{\infty} h_n(\theta_n) \leq f(\theta_0) - f^\star.$$

Thus, the nonnegative sequence $(h_n(\theta_n))_{n\geq 0}$ necessarily converges to zero. Then, we have two possibilities (according to the assumptions made in the proposition).

- If the functions g_n are majorizing f, we define $\theta' = \theta_n - \frac{1}{L}\nabla h_n(\theta_n)$, and we use the following classical inequality for L-smooth functions [44]:

  $$h_n(\theta') \leq h_n(\theta_n) - \frac{1}{2L}\|\nabla h_n(\theta_n)\|_2^2.$$

  Therefore, we may use the fact that $h_n(\theta') \geq 0$ because $g_n \geq f$, and

  $$\|\nabla h_n(\theta_n)\|_2^2 \leq 2L(h_n(\theta_n) - h_n(\theta')) \leq 2Lh_n(\theta_n) \underset{n\to+\infty}{\longrightarrow} 0.$$

- If instead the functions g_n are ρ-strongly convex, the last inequality of Lemma 2.3 with $\kappa = \theta = \theta_{n-1}$ and $\theta' = \theta_n$ gives us

  $$\frac{\rho}{2}\|\theta_n - \theta_{n-1}\|_2^2 \leq f(\theta_{n-1}) - f(\theta_n).$$

  By summing over n, we obtain that $\|\theta_n - \theta_{n-1}\|_2^2$ converges to zero, and

  $$\|\nabla h_n(\theta_n)\|_2 = \|\nabla h_n(\theta_n) - \nabla h_n(\theta_{n-1})\|_2 \leq L\|\theta_n - \theta_{n-1}\|_2 \underset{n\to+\infty}{\longrightarrow} 0,$$

  since $\nabla h_n(\theta_{n-1}) = 0$ according to Definition 2.2.

We now consider the directional derivative of f at θ_n and a direction $\theta - \theta_n$, where $n \geq 1$ and θ is in Θ,

$$\nabla f(\theta_n, \theta - \theta_n) = \nabla g_n(\theta_n, \theta - \theta_n) - \nabla h_n(\theta_n)^\top(\theta - \theta_n).$$

Note that θ_n minimizes g_n on Θ and therefore $\nabla g_n(\theta_n, \theta - \theta_n) \geq 0$. Therefore,

$$\nabla f(\theta_n, \theta - \theta_n) \geq -\|\nabla h_n(\theta_n)\|_2\|\theta - \theta_n\|_2$$

by the Cauchy–Schwarz inequality. By minimizing over θ and taking the infimum limit, we finally obtain

$$\liminf_{n\to+\infty} \inf_{\theta\in\Theta} \frac{\nabla f(\theta_n, \theta - \theta_n)}{\|\theta - \theta_n\|_2} \geq -\lim_{n\to+\infty} \|\nabla h_n(\theta_n)\|_2 = 0. \qquad (2.2)$$

∎

This proposition provides convergence guarantees for a large class of existing algorithms, including cases where f is nonsmooth. In the next proposition, we relax some of the assumptions for objective functions that are compositions $f = f' \circ e$, where $\circ$ is the composition operator. In other words, $f(\theta) = f'(e(\theta))$ for all θ in $\mathbb{R}^p$.

PROPOSITION 2.6 (nonconvex analysis for Algorithm 1—composition). *Assume that* (A) *holds and that the function f is a composition $f = f' \circ e$, where $e : \mathbb{R}^p \to \mathbb{R}^d$ is C-Lipschitz continuous for some constant $C > 0$, and $f' : \mathbb{R}^d \to \mathbb{R}$. Assume that the function g_n in Algorithm 1 is defined as $g_n \triangleq g'_n \circ e$, where g'_n is a majorizing surrogate in $\mathcal{S}_L(f', e(\theta_{n-1}))$. Then, the conclusions of Proposition 2.5 hold.*

Proof. We follow the same steps as the proof of Proposition 2.5. First, it is easy to show that $(f(\theta_n))_{n\geq 0}$ monotonically decreases and that $h_n(\theta_n) \triangleq g_n(\theta_n) - f(\theta_n)$ converges to zero when n grows to infinity. Note that since we have made the assumptions that $g_n = g'_n \circ e$ and that $f = f' \circ e$, the function $h_n \triangleq g_n - f$ can be written as $h_n = h'_n \circ e$, where $h'_n \triangleq g'_n - f'$ is L-smooth. Proceeding as in the proof of Proposition 2.5, we can show that $\|\nabla h'_n(e(\theta_n))\|_2$ converges to zero.

Let us now fix $n \geq 1$ and consider δ such that $\theta_n + \delta$ is in Θ. We have

$$h_n(\theta_n + \delta) = h'_n(e(\theta_n + \delta)) = h'_n(e(\theta_n) + \|\delta\|_2 \mathbf{z}),$$

where $\mathbf{z}$ is a vector whose ℓ_2-norm is bounded by a universal constant $C > 0$ because the function e is Lipschitz continuous. Since h'_n is L-smooth, we also have

$$h_n(\theta_n + \delta) = h'_n(e(\theta_n) + \|\delta\|_2 \mathbf{z}) = h_n(\theta_n) + \|\delta\|_2 \nabla h'_n(e(\theta_n))^\top \mathbf{z} + O(\|\delta\|_2^2).$$

Plugging this simple relation with $\delta = t(\theta - \theta_n)$, for some $0 < t < 1$ and θ in Θ, into the definition of the directional derivative $\nabla h_n(\theta_n, \theta - \theta_n)$, we obtain the relation

$$|\nabla h_n(\theta_n, \theta - \theta_n)| \leq C \|\nabla h'_n(e(\theta_n))\|_2 \|\theta - \theta_n\|_2,$$

and since $\nabla f(\theta_n, \theta - \theta_n) = \nabla g_n(\theta_n, \theta - \theta_n) - \nabla h_n(\theta_n, \theta - \theta_n)$, and $\nabla g_n(\theta_n, \theta - \theta_n) \geq 0$,

$$\liminf_{n \to +\infty} \inf_{\theta \in \Theta} \frac{\nabla f(\theta_n, \theta - \theta_n)}{\|\theta - \theta_n\|_2} \geq -C \lim_{n \to +\infty} \|\nabla h'_n(e(\theta_n))\|_2 = 0. \qquad \square$$

In this proposition, g_n is an upper bound of $f = f' \circ e$, where part e is Lipschitz continuous but $g_n - f$ is not L-smooth. This extension of Proposition 2.5 is useful since it provides convergence results for classical approaches that will be described later in section 2.3. Note that convergence results for nonconvex problems are by nature weak, and our nonconvex analysis does not provide any convergence rate. This is not the case when f is convex, as shown in the next section.

2.2. Convex analysis. The next proposition is based on a proof technique from Nesterov [45], which was originally designed for the proximal gradient method. By adapting it, we obtain the same convergence rates as in [45].

PROPOSITION 2.7 (convex analysis for $\mathcal{S}_L(f, \kappa)$). *Assume that f is convex and bounded below and that there exists a constant $R > 0$ such that*

$$\|\theta - \theta^\star\|_2 \leq R \quad \text{for all } \theta \in \Theta \ \text{ s.t. } \ f(\theta) \leq f(\theta_0), \tag{2.3}$$

where $\theta^\star$ is a minimizer of f on Θ. When the functions g_n in Algorithm 1 are in $\mathcal{S}_L(f, \theta_{n-1})$, we have for all $n \geq 1$,

$$f(\theta_n) - f^\star \leq \frac{2LR^2}{n+2},$$

where $f^\star \triangleq f(\theta^\star)$. Assume now that f is μ-strongly convex. Regardless of condition (2.3)*, we have for all $n \geq 1$,*

$$f(\theta_n) - f^\star \leq \beta^n (f(\theta_0) - f^\star),$$

where $\beta \triangleq \frac{L}{\mu}$ if $\mu > 2L$ or $\beta \triangleq \left(1 - \frac{\mu}{4L}\right)$ otherwise.

Proof. We successively prove the two parts of the proposition.

Nonstrongly convex case. Let us consider the function $h_n \triangleq g_n - f$ at iteration $n \geq 1$. By Lemma 2.3,

$$f(\theta_n) \leq \min_{\theta \in \Theta} \left[f(\theta) + \frac{L}{2} \|\theta - \theta_{n-1}\|_2^2 \right].$$

Then, following a similar proof technique as Nesterov in [45],

$$\begin{aligned} f(\theta_n) &\leq \min_{\alpha \in [0,1]} \left[f(\alpha\theta^\star + (1-\alpha)\theta_{n-1}) + \frac{L\alpha^2}{2} \|\theta^\star - \theta_{n-1}\|_2^2 \right] \\ &\leq \min_{\alpha \in [0,1]} \left[\alpha f(\theta^\star) + (1-\alpha) f(\theta_{n-1}) + \frac{L\alpha^2}{2} \|\theta^\star - \theta_{n-1}\|_2^2 \right], \end{aligned} \tag{2.4}$$

where the minimization over Θ is replaced by a minimization over the line segment $\alpha\theta^\star + (1-\alpha)\theta_{n-1} : \alpha \in [0,1]$. Since the sequence $(f(\theta_n))_{n \geq 0}$ is monotonically decreasing we may use the bounded level set assumption and we obtain

$$f(\theta_n) - f^\star \leq \min_{\alpha \in [0,1]} \left[(1-\alpha)(f(\theta_{n-1}) - f^\star) + \frac{LR^2\alpha^2}{2} \right]. \tag{2.5}$$

To simplify, we introduce the notation $r_n \triangleq f(\theta_n) - f^\star$, and we consider two cases:

- *First case:* If $r_{n-1} \geq LR^2$, then the optimal value $\alpha^\star$ in (2.5) is 1 and we consequently have $r_n \leq \frac{LR^2}{2}$.
- *Second case:* Otherwise $\alpha^\star = \frac{r_{n-1}}{LR^2}$ and $r_n \leq r_{n-1}\left(1 - \frac{r_{n-1}}{2LR^2}\right)$. Thus, $r_n^{-1} \geq r_{n-1}^{-1}\left(1 - \frac{r_{n-1}}{2LR^2}\right)^{-1} \geq r_{n-1}^{-1} + \frac{1}{2LR^2}$, where the second inequality comes from the convexity inequality $(1-x)^{-1} \geq 1 + x$ for $x \in (0,1)$.

We now apply recursively the previous inequalities, starting with $n = 1$. If $r_0 \geq LR^2$, we are in the first case and then $r_1 \leq \frac{LR^2}{2}$. Then, we will subsequently be in the second case for all $n \geq 2$ and thus $r_n^{-1} \geq r_1^{-1} + \frac{n-1}{2LR^2} \geq \frac{n+3}{2LR^2}$. Otherwise, if $r_0 < LR^2$, we are always in the second case and $r_n^{-1} \geq r_0^{-1} + \frac{n}{2LR^2} \geq \frac{n+2}{2LR^2}$, which is sufficient to obtain the first part of the proposition.

μ-strongly convex case. Let us now assume that f is μ-strongly convex, and let us drop the bounded level sets assumption. The proof again follows [45] for computing the convergence rate of proximal gradient methods. We start from (2.4). We use the strong convexity of f which implies that $f(\theta_{n-1}) \geq f^\star + \frac{\mu}{2}\|\theta_{n-1} - \theta^\star\|_2^2$, and we obtain

$$f(\theta_n) - f^\star \leq \left(\min_{\alpha \in [0,1]} 1 - \alpha + \frac{L\alpha^2}{\mu} \right) (f(\theta_{n-1}) - f^\star).$$

At this point, it is easy to show that if $\mu \geq 2L$, the previous binomial is minimized for $\alpha^\star = 1$, and if $\mu \leq 2L$, then we have $\alpha^\star = \frac{\mu}{2L}$. This yields the desired result. ∎

The result of Proposition 2.7 is interesting because it does not make any strong assumption about the surrogate functions, except the ones from Definition 2.2. The next proposition shows that slightly better rates can be obtained with additional strong convexity assumptions.

PROPOSITION 2.8 (convex analysis for $\mathcal{S}_{L,\rho}(f,\kappa)$). *Assume that f is convex and bounded below, and let $\theta^\star$ be a minimizer of f on Θ. When the surrogates g_n of Algorithm 1 are in $\mathcal{S}_{L,\rho}(f,\theta_{n-1})$ with $\rho \geq L$, we have for all $n \geq 1$,*

$$f(\theta_n) - f^\star \leq \frac{L\|\theta_0 - \theta^\star\|_2^2}{2n},$$

where $f^\star \triangleq f(\theta^\star)$. When f is μ-strongly convex, we have for all $n \geq 1$,

$$f(\theta_n) - f^\star \leq \left(\frac{L}{\rho+\mu}\right)^{n-1} \frac{L\|\theta_0 - \theta^\star\|_2^2}{2}.$$

Proof. As before, we successively prove the two parts of the proposition.

Nonstrongly convex case. From Lemma 2.3 (with $g = g_n$, $\kappa = \theta_{n-1}$, $\theta' = \theta_n$, $\theta = \theta^\star$), we have for all $n \geq 1$,

$$f(\theta_n) - f(\theta^\star) \leq \frac{L}{2}\|\theta_{n-1} - \theta^\star\|_2^2 - \frac{\rho}{2}\|\theta_n - \theta^\star\|_2^2 \leq \frac{L}{2}\|\theta_{n-1} - \theta^\star\|_2^2 - \frac{L}{2}\|\theta_n - \theta^\star\|_2^2. \tag{2.6}$$

After summation,

$$n(f(\theta_n) - f(\theta^\star)) \leq \sum_{k=1}^{n}(f(\theta_k) - f(\theta^\star)) \leq \frac{L}{2}(\|\theta_0 - \theta^\star\|_2^2 - \|\theta_n - \theta^\star\|_2^2) \leq \frac{L\|\theta_0 - \theta^\star\|_2^2}{2},$$

where the first inequality comes from the inequalities $f(\theta_k) \geq f(\theta_n)$ for all $k \leq n$. This is sufficient to prove the first part. Note that proving convergence rates for first-order methods by finding telescopic sums is a classical technique (see, e.g., [3]).

μ-strongly convex case. Let us now assume that f is μ-strongly convex. The strong convexity implies that $f(\theta_n) - f^\star \geq \frac{\mu}{2}\|\theta_n - \theta^\star\|_2^2$ for all n. Combined with (2.6), this yields

$$\frac{\mu+\rho}{2}\|\theta_n - \theta^\star\|_2^2 \leq \frac{L}{2}\|\theta_{n-1} - \theta^\star\|_2^2,$$

and thus

$$f(\theta_n) - f(\theta^\star) \leq \frac{L}{2}\|\theta_{n-1} - \theta^\star\|_2^2 \leq \left(\frac{L}{\rho+\mu}\right)^{n-1} \frac{L\|\theta_0 - \theta^\star\|_2^2}{2}. \qquad \square$$

Even though the constants obtained in the rates of Proposition 2.8 are slightly better than the ones of Proposition 2.7, the condition g_n in $\mathcal{S}_{L,\rho}(f,\kappa)$ with $\rho \geq L$ is much stronger than the simple assumption that g_n is in $\mathcal{S}_L(f,\kappa)$. It can indeed be shown that f is necessarily $(\rho - L)$-strongly convex if $\rho > L$ and convex if $\rho = L$. In the next section, we give some examples where such a condition holds.

2.3. Examples of first-order surrogate functions. We now present practical first-order surrogate functions and links between Algorithm 1 and existing approaches. Even though our generic analysis does not always bring new results for each specific case, its main asset is to provide a unique theoretical treatment to all of them.

2.3.1. Lipschitz gradient surrogates. When f is L-smooth, it is natural to consider the following surrogate:

$$g : \theta \mapsto f(\kappa) + \nabla f(\kappa)^\top (\theta - \kappa) + \frac{L}{2}\|\theta - \kappa\|_2^2.$$

The function g is an upper bound of f, which is a classical result [44]. It is then easy to see that g is L-strongly convex and L-smooth. As a consequence, the difference $g - f$ is $2L$-smooth (as a sum of two L-smooth functions), and thus g is in $\mathcal{S}_{2L,L}(f, \kappa)$.

When f is convex, it is also possible to show by using Lemma A.5 that g is in fact in $\mathcal{S}_{L,L}(f, \kappa)$, and when f is μ-strongly convex, g is in $\mathcal{S}_{L-\mu,L}(f, \kappa)$. We remark that minimizing g amounts to performing a gradient descent step: $\theta' \leftarrow \kappa - \frac{1}{L}\nabla f(\kappa)$.

2.3.2. Proximal gradient surrogates. Let us now consider a composite optimization problem, meaning that f splits into two parts $f = f_1 + f_2$, where f_1 is L-smooth. Then, a natural surrogate of f is the following function:

$$g : \theta \mapsto f_1(\kappa) + \nabla f_1(\kappa)^\top (\theta - \kappa) + \frac{L}{2}\|\theta - \kappa\|_2^2 + f_2(\theta).$$

The function g majorizes f and the approximation error $g - f$ is the same as in section 2.3.1. Thus, g in in $\mathcal{S}_{2L}(f, \kappa)$ or in $\mathcal{S}_{2L,L}(f, \kappa)$ when f_2 is convex. Moreover,

- when f_1 is convex, g is in $\mathcal{S}_L(f, \kappa)$, and if f_2 is also convex, g is in $\mathcal{S}_{L,L}(f, \kappa)$;
- when f_1 is μ-strongly convex, g is in $\mathcal{S}_{L-\mu}(f, \kappa)$. If f_2 is also convex, g is in $\mathcal{S}_{L-\mu,L}(f, \kappa)$.

Minimizing g amounts to performing one step of the proximal gradient algorithm [3, 45, 54]. It is indeed easy to show that the minimum θ' of g—assuming it is unique—can be equivalently obtained as follows:

$$\theta' = \operatorname*{arg\,min}_{\theta \in \Theta} \left[\frac{1}{2} \left\| \theta - \left(\kappa - \frac{1}{L}\nabla f_1(\kappa) \right) \right\|_2^2 + \frac{1}{L} f_2(\theta) \right],$$

which is often written under the form $\theta' = \mathrm{Prox}_{f_2/L}[\kappa - (1/L)\nabla f_1(\kappa)]$, where "Prox" is called the "proximal operator" [41]. In some cases, the proximal operator can be computed efficiently in closed form, for example, when f_2 is the ℓ_1-norm; it yields the iterative soft-thresholding algorithm for sparse estimation [16]. For a review of proximal operators and their computations, we refer the reader to [2, 15].

2.3.3. Linearizing concave functions and DC programming. Assume that $f = f_1 + f_2$, where f_2 is concave and L-smooth. Then, the following function g is a majorizing surrogate in $\mathcal{S}_L(f, \kappa)$:

$$g : \theta \mapsto f_1(\theta) + f_2(\kappa) + \nabla f_2(\kappa)^\top (\theta - \kappa).$$

Such a surrogate appears in DC programming [30]. When f_1 is convex, f is indeed the difference of two convex functions. It is also used in sparse estimation for dealing with some nonconvex penalties [2]. For example, consider a cost function of the form $\theta \mapsto f_1(\theta) + \lambda \sum_{j=1}^p \log(|\theta[j]| + \varepsilon)$, where $\theta[j]$ is the jth entry in θ. Even though the functions $\theta \mapsto \log(|\theta[j]| + \varepsilon)$ are not differentiable, they can be written as the composition of a concave smooth function $u \mapsto \log(u + \varepsilon)$ on $\mathbb{R}^+$ and a Lipschitz function $\theta \mapsto |\theta[j]|$. By upper-bounding the logarithm function by its linear approximation, Proposition 2.6 justifies the surrogate

$$g : \theta \mapsto f_1(\theta) + \lambda \sum_{j=1}^p \log(|\kappa[j]| + \varepsilon) + \lambda \sum_{j=1}^p \frac{|\theta[j]| - |\kappa[j]|}{|\kappa[j]| + \varepsilon}, \tag{2.7}$$

and minimizing g amounts to performing one step of a reweighted-ℓ_1 algorithm (see [11] and references therein). Similarly, other penalty functions are adapted to this framework. For instance, the logarithm can be replaced by any smooth concave nondecreasing function, or group-sparsity penalties [52, 56] can be used, such as $\theta \mapsto \sum_{g\in\mathcal{G}} \log(\|\theta_g\|_2 + \varepsilon)$, where $\mathcal{G}$ is a partition of $\{1,\dots,p\}$ and θ_g records the entries of θ corresponding to the set g. Proposition 2.6 indeed applies to this setting.

2.3.4. Variational surrogates. Let us now consider a real-valued function f defined on $\mathbb{R}^{p_1}\times\mathbb{R}^{p_2}$. Let $\Theta_1 \subseteq \mathbb{R}^{p_1}$ and $\Theta_2 \subseteq \mathbb{R}^{p_2}$ be two convex sets. Minimizing f over $\Theta_1\times\Theta_2$ is equivalent to minimizing the function $\tilde{f}$ over Θ_1 defined as $\tilde{f}(\theta_1) \triangleq\mapsto \min_{\theta_2\in\Theta_2} f(\theta_1,\theta_2)$. Assume now that

- $\theta_2 \mapsto f(\theta_1,\theta_2)$ is μ-strongly convex for all θ_1 in $\mathbb{R}^{p_1}$;
- $\theta_1 \mapsto f(\theta_1,\theta_2)$ is differentiable for all θ_2;
- $(\theta_1,\theta_2) \mapsto \nabla_1 f(\theta_1,\theta_2)$ is L'-Lipschitz with respect to θ_1 and L-Lipschitz with respect to θ_2.[1]

Let us fix κ_1 in Θ_1. Then, the following function is a majorizing surrogate in $\mathcal{S}_{L''}(\tilde{f},\kappa)$:

$$g : \theta_1 \mapsto f(\theta_1,\kappa_2^\star) \text{ with } \kappa_2^\star \triangleq \operatorname*{arg\,min}_{\theta_2\in\Theta_2} f(\kappa_1,\theta_2)$$

with $L'' = 2L' + L^2/\mu$. We can indeed apply Lemma A.6, which ensures that $\tilde{f}$ is differentiable with $\nabla\tilde{f}(\theta_1) = \nabla_1 f(\theta_1,\theta_2^\star)$ and $\theta_2^\star \triangleq \arg\min f(\theta_1,\theta_2)$ for all θ_1. Moreover, g is L'-smooth and $\tilde{f}$ is $L' + L^2/\mu$-smooth according to Lemma A.6, and thus $h \triangleq g - \tilde{f}$ is L''-smooth. Note that a better constant $L'' = L'$ can be obtained when f is convex, as noted in the appendix of [38].

The surrogate g leads to an alternate minimization algorithm; it is then interesting to note that Proposition 2.7 provides similar convergence rates as another recent analysis [4], which makes slightly different assumptions on the function f. Variational surrogates might also be useful for problems of a single variable θ_1. For instance, consider a regression problem with a Huber loss function H defined for all u in $\mathbb{R}$ as

$$H(u) \triangleq \begin{cases} \frac{u^2}{2\delta} + \frac{\delta}{2} & \text{if } |u| \le \delta, \\ |u| & \text{otherwise,} \end{cases}$$

where δ is a positive constant.[2] The Huber loss can be seen as a smoothed version of the ℓ_1-norm when δ is small, or simply a robust variant of the squared loss $u \mapsto \frac{1}{2}u^2$ that asymptotically grows linearly. Then, it is easy to show that

$$H(u) = \frac{1}{2}\min_{w\ge\delta}\left[\frac{u^2}{w} + w\right].$$

Consider now a regression problem with m training data points represented by vectors $\mathbf{x}_i$ in $\mathbb{R}^p$, associated to real numbers y_i, for $i = 1,\dots,m$. The robust regression problem with the Huber loss can be formulated as the minimization over $\mathbb{R}^p$ of

$$\tilde{f} : \theta_1 \mapsto \sum_{i=1}^m H(y_i - \mathbf{x}_i^\top\theta_1) = \min_{\theta_2\in\mathbb{R}^m:\theta_2\ge\delta}\left[f(\theta_1,\theta_2) \triangleq \frac{1}{2}\sum_{i=1}^m \frac{(y_i - \mathbf{x}_i^\top\theta_1)^2}{\theta_2[i]} + \theta_2[i]\right],$$

[1] The notation ∇_1 denotes the gradient with respect to θ_1.

[2] To simplify the notation, we present a shifted version of the traditional Huber loss, which usually satisfies $H(0) = 0$.

where θ_1 is the parameter vector of a linear model. The conditions described at the beginning of this section can be shown to be satisfied with a Lipschitz constant proportional to $(1/\delta)$; the resulting algorithm is the iterative reweighted least-square method, which appears in the literature about both robust statistics [35] and sparse estimation where the Huber loss is used to approximate the ℓ_1-norm [2].

2.3.5. Jensen surrogates. Jensen's inequality also provides a natural mechanism to obtain surrogates for convex functions. Following the presentation of Lange, Hunter, and Yang [35], we consider a convex function $f : \mathbb{R} \mapsto \mathbb{R}$ and a vector $\mathbf{x}$ in $\mathbb{R}^p$ and define $\tilde{f} : \mathbb{R}^p \to \mathbb{R}$ as $\tilde{f}(\theta) \triangleq f(\mathbf{x}^\top \theta)$ for all θ. Let $\mathbf{w}$ be a weight vector in $\mathbb{R}^p_+$ such that $\|\mathbf{w}\|_1 = 1$ and $\mathbf{w}[i] \neq 0$ whenever $\mathbf{x}[i] \neq 0$. Then, we define for any κ in $\mathbb{R}^p$,

$$g : \theta \mapsto \sum_{i=1}^{p} \mathbf{w}[i] f\left(\frac{\mathbf{x}[i]}{\mathbf{w}[i]}(\theta[i] - \kappa[i]) + \mathbf{x}^\top \kappa\right).$$

When f is L-smooth, and when $\mathbf{w}[i] \triangleq |\mathbf{x}[i]|^\nu / \|\mathbf{x}\|_\nu^\nu$, g is in $\mathcal{S}_{L'}(\tilde{f}, \kappa)$ with

- $L' = L\|\mathbf{x}\|_\infty^2 \|\mathbf{x}\|_0$ for $\nu = 0$;
- $L' = L\|\mathbf{x}\|_\infty \|\mathbf{x}\|_1$ for $\nu = 1$;
- $L' = L\|\mathbf{x}\|_2^2$ for $\nu = 2$.

To the best of our knowledge, nonasymptotic convergence rates have not been studied before for such surrogates, and thus we believe that our analysis may provide new results in the present case. Jensen surrogates are indeed quite uncommon; they appear nevertheless in a few occasions. In addition to the few examples given in [35], they are used, for instance, in machine learning in [19] for interpreting boosting procedures through the concept of *auxiliary functions*.

Jensen's inequality is also used in a different fashion in EM algorithms [20, 42]. Consider T nonnegative functions $f^t : \mathbb{R}^p \mapsto \mathbb{R}_+$, and, for some κ in $\mathbb{R}^p$, define some weights $\mathbf{w}[t] = f^t(\kappa) / \sum_{t'=1}^{T} f^{t'}(\kappa)$. By exploiting the concavity of the logarithm, and assuming that $\mathbf{w}[t] > 0$ for all t to simplify, Jensen's inequality yields

$$-\log\left(\sum_{t=1}^{T} f^t(\theta)\right) \leq -\sum_{t=1}^{T} \mathbf{w}[t] \log\left(\frac{f^t(\theta)}{\mathbf{w}[t]}\right). \tag{2.8}$$

The relation (2.8) is key to EM algorithms minimizing a negative log-likelihood. The right side of this equation can be interpreted as a majorizing surrogate of the left side since it is easy to show that both terms are equal for $\theta = \kappa$. Unfortunately the resulting approximation error functions are not L-smooth in general and these surrogates do not follow the assumptions of Definition 2.2. As a consequence, our analysis may apply to some EM algorithms, but not to all of them.

2.3.6. Quadratic surrogates. When f is twice differentiable and admits a matrix $\mathbf{H}$ such that $\mathbf{H} - \nabla^2 f$ is always positive definite, the following function is a first-order majorizing surrogate:

$$g : \theta \mapsto f(\kappa) + \nabla f(\kappa)^\top (\theta - \kappa) + \frac{1}{2}(\theta - \kappa)^\top \mathbf{H}(\theta - \kappa).$$

The Lipschitz constant of $\nabla(g - f)$ is the largest eigenvalue of $\mathbf{H} - \nabla^2 f(\theta)$ over Θ. Such surrogates appear frequently in the statistics and machine learning literature [7, 31, 33]. The goal is to model the global curvature of the objective function during each iteration, without resorting to the Newton method. Even though quadratic surrogates do not necessarily lead to better theoretical convergence rates than simpler Lipschitz gradient surrogates, they can be quite effective in practice [31].

3. An incremental majorization-minimization algorithm: MISO. In this section, we introduce an incremental scheme that exploits the structure (1.1) of f as a large sum of T components. The most popular method for dealing with such a problem when f is smooth and $\Theta = \mathbb{R}^p$ is probably the *stochastic gradient descent* algorithm (SGD) and its variants (see [43]). It consists of drawing at iteration n an index $\hat{t}_n$ and updating the solution as $\theta_n \leftarrow \theta_{n-1} - \eta_n \nabla f^{\hat{t}_n}(\theta_{n-1})$, where the scalar η_n is a step size. Another popular algorithm is the *stochastic mirror descent* algorithm (see [32]) for general nonsmooth convex problems, a setting we do not consider in this paper since nonsmooth functions do not always admit practical first-order surrogates.

Recently, linear convergence rates for strongly convex functions f^t have been obtained in [50] and [51] by using randomized incremental algorithms whose cost per iteration is independent of T. The method SAG [50] for smooth unconstrained convex optimization is a randomized variant of the incremental gradient descent algorithm of Blatt, Hero, and Gauchman [6], where an estimate of the gradient ∇f is incrementally updated at each iteration. The method SDCA [51] for strongly convex composite optimization is a dual coordinate ascent algorithm that performs incremental updates in the primal (1.1). Unlike SGD, both SAG and SDCA require storing information about past iterates, which is a key for obtaining fast convergence rates.

In a different context, incremental EM algorithms have been proposed by Neal and Hinton [42], where upper bounds of a nonconvex negative log-likelihood function are incrementally updated. By using similar ideas, we introduce the scheme MISO in Algorithm 2. At every iteration, a single function is observed, and an approximate surrogate of f is updated. Note that in the same line of work, Ahn et al. [1] have proposed a block-coordinate descent majorization-minimization algorithm, which corresponds to MISO when the variational surrogates of section 2.3.4 are used.

Algorithm 2. Incremental scheme MISO.

input $\theta_0 \in \Theta$ (initial estimate); N (number of iterations).

1: Initialization: choose some surrogates g_0^t of f^t near θ_0 for all t;
2: **for** $n = 1, \ldots, N$ **do**
3: Randomly pick up one index $\hat{t}_n$ and choose a surrogate $g_n^{\hat{t}_n}$ of $f^{\hat{t}_n}$ near θ_{n-1}; set $g_n^t \triangleq g_{n-1}^t$ for all $t \neq \hat{t}_n$.
4: Update the solution: $\theta_n \in \arg\min_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T} g_n^t(\theta)$.
5: **end for**

output θ_N (final estimate);

In the next two sections, we study the convergence properties of the scheme MISO. We proceed as in section 2. Specifically, we start with the nonconvex case, focusing on stationary point conditions, and we show that similar guarantees as for the batch majorization-minimization algorithm hold. Then, for convex problems, we present convergence rates that essentially apply to the proximal gradient surrogates. We obtain sublinear rates $O(T/n)$ for the general convex case and linear ones for strongly convex objective functions. Even though these rates do not show any theoretical advantage over the batch algorithm, we also present a more surprising result in section 3.2; in a large sample regime $T \geq 2L/\mu$, for μ-strongly convex functions f^t, minorizing surrogates may be used and faster rates can be achieved.

3.1. Convergence analysis. We start our analysis with the nonconvex case and make the following assumption:

(B) f is bounded below and for all θ, θ' in Θ and all t, the directional derivative $\nabla f^t(\theta, \theta' - \theta)$ of f^t at θ in the direction $\theta' - \theta$ exists.

Then, we obtain a first convergence result.

PROPOSITION 3.1 (nonconvex analysis). *Assume that* (B) *holds and that the surrogates $g_n^{\hat{t}_n}$ from Algorithm* 2 *are majorizing $f^{\hat{t}_n}$ and are in $\mathcal{S}_L(f^{\hat{t}_n}, \theta_{n-1})$. Then, the conclusions of Proposition* 2.5 *hold with probability one.*

Proof. We proceed in several steps.

Almost sure convergence of $(f(\theta_n))_{n\geq 0}$. Let us define $\bar{g}_n \triangleq \frac{1}{T}\sum_{t=1}^T g_n^t$. We have the following relation for all $n \geq 1$:

$$\bar{g}_n = \bar{g}_{n-1} + \frac{g_n^{\hat{t}_n} - g_{n-1}^{\hat{t}_n}}{T}, \tag{3.1}$$

where the surrogates and the index $\hat{t}_n$ are chosen in the algorithm. Then, we obtain the following inequalities, which hold with probability one for all $n \geq 1$,

$$\begin{aligned}\bar{g}_n(\theta_n) &\leq \bar{g}_n(\theta_{n-1}) = \bar{g}_{n-1}(\theta_{n-1}) + \frac{g_n^{\hat{t}_n}(\theta_{n-1}) - g_{n-1}^{\hat{t}_n}(\theta_{n-1})}{T} \\ &= \bar{g}_{n-1}(\theta_{n-1}) + \frac{f^{\hat{t}_n}(\theta_{n-1}) - g_{n-1}^{\hat{t}_n}(\theta_{n-1})}{T} \leq \bar{g}_{n-1}(\theta_{n-1}).\end{aligned}$$

The first inequality is true by definition of θ_n and the second one because $\bar{g}_{n-1}^{\hat{t}_n}$ is a majorizing surrogate of $f^{\hat{t}_n}$. The sequence $(\bar{g}_n(\theta_n))_{n\geq 0}$ is thus monotonically decreasing, bounded below with probability one, and thus converges almost surely. By taking the expectation of these previous inequalities, we also obtain that the sequence $(\mathbb{E}[\bar{g}_n(\theta_n)])_{n\geq 0}$ monotonically converges. Thus, the nonpositive quantity $\mathbb{E}[f^{\hat{t}_n}(\theta_{n-1}) - g_{n-1}^{\hat{t}_n}(\theta_{n-1})]$ is the summand of a converging sum and we have

$$\begin{aligned}\mathbb{E}\left[\sum_{n=0}^{+\infty} g_n^{\hat{t}_{n+1}}(\theta_n) - f^{\hat{t}_{n+1}}(\theta_n)\right] &= \sum_{n=0}^{+\infty} \mathbb{E}[g_n^{\hat{t}_{n+1}}(\theta_n) - f^{\hat{t}_{n+1}}(\theta_n)] \\ &= \sum_{n=0}^{+\infty} \mathbb{E}[\mathbb{E}[g_n^{\hat{t}_{n+1}}(\theta_n) - f^{\hat{t}_{n+1}}(\theta_n)|\mathcal{F}_n]] \\ &= \sum_{n=0}^{+\infty} \mathbb{E}[\bar{g}_n(\theta_n) - f(\theta_n)] \\ &= \mathbb{E}\left[\sum_{n=0}^{+\infty} \bar{g}_n(\theta_n) - f(\theta_n)\right] < +\infty,\end{aligned}$$

where we use the Beppo–Lévy theorem to interchange the expectation and the sum in front of nonnegative quantities, and $\mathcal{F}_n$ is the filtration representing all information up to iteration n (including θ_n). As a result, the sequence $(\bar{g}_n(\theta_n) - f(\theta_n))_{n\geq 0}$ converges almost surely to 0, implying the almost sure convergence of $(f(\theta_n))_{n\geq 0}$.

Asymptotic stationary point conditions. Let us define $\bar{h}_n \triangleq \bar{g}_n - f$, which is L-smooth. Then, for all θ in Θ and $n \geq 1$,

$$\nabla f(\theta_n, \theta - \theta_n) = \nabla \bar{g}_n(\theta_n, \theta - \theta_n) - \nabla \bar{h}_n(\theta_n)^\top(\theta - \theta_n).$$

We have $\nabla \bar{g}_n(\theta_n, \theta - \theta_n) \geq 0$ by definition of θ_n, and $\|\nabla \bar{h}_n(\theta_n)\|_2^2 \leq 2L\bar{h}_n(\theta_n)$, following similar steps as in the proof of Proposition 2.5. Since we have previously

shown that $(\bar{h}_n(\theta_n))_{n\geq 0}$ almost surely converges to zero, we conclude as in the proof of Proposition 2.5, replacing h_n by $\bar{h}_n$ and g_n by $\bar{g}_n$. ∎

We also give the counterpart of Proposition 2.6 for Algorithm 2.

PROPOSITION 3.2 (nonconvex analysis—composition). *Assume that* (B) *is satisfied and that the functions f^t are compositions $f^t = f'^t \circ e^t$, where the functions e^t are C-Lipschitz continuous for some $C > 0$. Assume also that the functions $g_n^{\hat{t}_n}$ in Algorithm* 2 *are also compositions $g_n^{\hat{t}_n} = g_n'^{\hat{t}_n} \circ e^{\hat{t}_n}$, where $g_n'^{\hat{t}_n}$ is majorizing $f'^{\hat{t}_n}$ and is in $\mathcal{S}_L(f'^{\hat{t}_n}, e^{\hat{t}_n}(\theta_{n-1}))$. Then, the conclusions of Proposition* 3.1 *hold.*

Proof. We first remark that the first part of the proof of Proposition 3.1 does not exploit the fact that the approximation errors $g_n^t - f^t$ are L-smooth, but only the fact that g_n^t is majorizing f^t for all n and t. Thus, the first part of the proof of Proposition 3.1 holds in the present case, such that $(f(\theta_n))_{n\geq 0}$ almost surely converges, and the sequence $(\bar{g}_n(\theta_n) - f(\theta_n))_{n\geq 0}$ almost surely converges to zero, where $\bar{g}_n$ is defined in the proof of Proposition 3.1.

It remains to show that the asymptotic stationary point conditions are satisfied. To that effect, we follow the proof of Proposition 2.6. We first have, for all $n \geq 1$,

$$\nabla f(\theta_n, \theta - \theta_n) = \nabla \bar{g}_n(\theta_n, \theta - \theta_n) - \frac{1}{T}\sum_{t=1}^{T} \nabla \bar{h}_n^t(\theta_n, \theta - \theta_n),$$

with $\nabla \bar{g}_n(\theta_n, \theta - \theta_n) \geq 0$ and $\bar{h}_n^t \triangleq \bar{g}_n^t - f^t$. Then, following the proof of Proposition 2.6, it is easy to show that

$$|\nabla \bar{h}_n^t(\theta_n, \theta - \theta_n)| \leq C \|\nabla \bar{h}_n'^t(e^t(\theta_n))\|_2 \|\theta - \theta_n\|_2,$$

where $\bar{h}_n'^t = \bar{g}_n'^t - f'^t$, and we conclude as in Proposition 2.6. ∎

The next lemma provides convergence rates for the convex case, under the assumption that the surrogate functions are ρ-strongly convex with $\rho \geq L$. The result notably applies to the proximal gradient surrogates of section 2.3.2.

PROPOSITION 3.3 (convex analysis for strongly convex surrogate functions). *Assume that f is convex and bounded below, let $\theta^\star$ be a minimizer of f on Θ, and let us define $f^\star \triangleq \min_{\theta\in\Theta} f(\theta)$. When the surrogates g_n^t in Algorithm* 2 *are majorizing f^t and are in $\mathcal{S}_{L,\rho}(f^t, \theta_{n-1})$ with $\rho \geq L$, we have for all $n \geq 1$,*

$$\mathbb{E}[f(\bar{\theta}_n) - f^\star] \leq \frac{LT\|\theta^\star - \theta_0\|_2^2}{2n}, \tag{3.2}$$

where $\bar{\theta}_n \triangleq \frac{1}{n}\sum_{i=1}^n \theta_i$ is the average of the iterates. Assume now that f is μ-strongly convex. For all $n \geq 1$,

$$\mathbb{E}[f(\theta_n) - f^\star] \leq \left(1 - \frac{2\mu}{T(\rho+\mu)}\right)^{n-1} \frac{L\|\theta^\star - \theta_0\|_2^2}{2}. \tag{3.3}$$

Proof. We proceed in several steps.

Preliminaries. For all $n \geq 1$, we introduce the point κ_{n-1}^t in Θ such that g_n^t is in $\mathcal{S}_{L,\rho}(f^t, \kappa_{n-1}^t)$. We remark that such points are drawn recursively according to the following conditional probability distribution:

$$\mathbb{P}(\kappa_{n-1}^t = \theta_{n-1}|\mathcal{F}_{n-1}) = \delta \quad \text{and} \quad \mathbb{P}(\kappa_{n-1}^t = \kappa_{n-2}^t|\mathcal{F}_{n-1}) = 1 - \delta,$$

where $\delta \triangleq 1/T$, $\mathcal{F}_n$ is the filtration representing all information up to iteration n (including θ_n), and $\kappa_0^t \triangleq \theta_0$ for all t. Thus we have for all t and all $n \geq 1$,

(3.4)
$$\mathbb{E}[\|\theta^\star - \kappa_{n-1}^t\|_2^2] = \mathbb{E}[\mathbb{E}[\|\theta^\star - \kappa_{n-1}^t\|_2^2 | \mathcal{F}_{n-1}]] = \delta \mathbb{E}[\|\theta^\star - \theta_{n-1}\|_2^2] + (1-\delta)\mathbb{E}[\|\theta^\star - \kappa_{n-2}^t\|_2^2].$$

We also need the following extension of Lemma 2.3 to the incremental setting: for all θ in Θ and $n \geq 1$,

$$f(\theta_n) \leq f(\theta) + \frac{1}{T}\sum_{t=1}^{T}\left(\frac{L}{2}\|\theta - \kappa_{n-1}^t\|_2^2 - \frac{\rho}{2}\|\theta - \theta_n\|_2^2\right). \tag{3.5}$$

The proof of this relation is similar to that of Lemma 2.3, exploiting the ρ-strong convexity of $\bar{g}_n \triangleq (1/T)\sum_{t=1}^T g_n^t$. We can now study the first part of the proposition.

Nonstrongly convex case ($\rho = L$). Let us define the quantities $A_n \triangleq \mathbb{E}[\frac{1}{2T}\sum_{t=1}^T \|\theta^\star - \kappa_n^t\|_2^2]$ and $\xi_n \triangleq \frac{1}{2}\mathbb{E}[\|\theta^\star - \theta_n\|_2^2]$. Then, we have from (3.5) with $\theta = \theta^\star$, and by taking the expectation

$$\mathbb{E}[f(\theta_n) - f^\star] \leq LA_{n-1} - L\xi_n.$$

It follows from (3.4) that $A_n = \delta\xi_n + (1-\delta)A_{n-1}$ and thus, for all $n \geq 1$,

$$\mathbb{E}[f(\theta_n) - f^\star] \leq \frac{L}{\delta}(A_{n-1} - A_n).$$

By summing the above inequalities, and using Jensen's inequality, we obtain that

$$\mathbb{E}[f(\bar{\theta}_n) - f^\star] \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[f(\theta_i) - f^\star] \leq \frac{LA_0}{\delta},$$

leading to the convergence rate of Eq. (3.2), since $A_0 = \frac{1}{2}\|\theta^\star - \theta_0\|_2^2$.

μ-strongly convex case. Assume now that the functions f^t are μ-strongly convex. For all $n \geq 1$, the strong convexity of f and (3.5) give us the following inequalities:

$$\mu\xi_n \leq \mathbb{E}[f(\theta_n) - f^\star] \leq LA_{n-1} - \rho\xi_n.$$

Combining this last inequality with (3.4), we obtain that for all $n \geq 1$,

$$A_n = \delta\xi_n + (1-\delta)A_{n-1} \leq \left(\frac{\delta L}{\mu + \rho} + (1-\delta)\right)A_{n-1}.$$

Thus, $A_n \leq \beta^n A_0$ with $\beta \triangleq \frac{(1-\delta)(\rho+\mu)+\delta L}{\rho+\mu}$. Since $A_0 = \xi_0$, $\mathbb{E}[f(\theta_n) - f^\star] \leq LA_{n-1}$, and $\beta \leq 1 - 2\delta\mu/(\rho+\mu)$, we finally have shown the desired convergence rate (3.3). ∎

The convergence rate of the previous proposition in the convex case suggests that the incremental scheme and batch one of section 2 have the same overall complexity, assuming that each iteration of the batch algorithm is T times the one of MISO. For strongly convex functions f^t, we obtain linear convergence rates, a property shared by SAG or SDCA; it is thus natural to make a more precise comparison with these other incremental approaches, which we present in the next two sections.

3.2. MISO for smooth unconstrained optimization. In this section, we assume that the optimization domain is unbounded—that is, $\Theta = \mathbb{R}^p$—and that the functions f^t are L-smooth. When using the Lipschitz gradient surrogates of section 2.3.1, MISO amounts to iteratively using the following update rule:

$$\theta_n \leftarrow \frac{1}{T}\sum_{t=1}^{T} \kappa_{n-1}^t - \frac{1}{LT}\sum_{t=1}^{T} \nabla f^t(\kappa_{n-1}^t), \tag{3.6}$$

where the vectors κ_{n-1} are recursively defined for $n \geq 2$ as $\kappa_{n-1}^{\hat{t}_n} = \theta_{n-1}$ and $\kappa_{n-1}^t = \kappa_{n-2}^t$ for $t \neq \hat{t}_n$, with $\kappa_0^t = \theta_0$ for all t. It is then easy to see that the complexity of updating θ_n is independent of T by storing the vectors $\mathbf{z}_n^t = \kappa_{n-1}^t - (1/L)\nabla f^t(\kappa_{n-1}^t)$ and performing the update $\theta_n = \theta_{n-1} + (1/T)(\mathbf{z}_n^t - \mathbf{z}_{n-1}^t)$. In comparison, the approach SAG yields a different, but related, update rule:

$$\theta_n \leftarrow \theta_{n-1} - \frac{\alpha}{T}\sum_{t=1}^{T} \nabla f^t(\kappa_{n-1}^t), \tag{3.7}$$

where the value $\alpha = 1/(16L)$ is suggested in [50]. Even though the rules (3.6) and (3.7) seem to be similar to each other at first sight, they behave differently in practice and do not have the same theoretical properties. For nonconvex problems, MISO is guaranteed to converge, whereas it is not known whether this is the case for SAG. For convex problems, both methods have a convergence rate of the same nature—that is, $O(T/n)$. For μ-strongly-convex problems, however, the convergence rate of SAG reported in [50] is substantially better than ours. Whereas the expected objective of SAG decreases with the rate $O(\rho^n)$ with $\rho_{\text{SAG}} = 1 - \min(\mu/(16L), 1/(8T))$, ours decreases with $\rho_{\text{MISO}} = 1 - 2\mu/(T(L+\mu))$, which is larger than ρ_{SAG} unless the problem is very well conditioned.

By maximizing the convex dual of (1.1) when the functions f^t are μ-strongly convex, the approach SDCA yields another update rule that resembles (3.6) and (3.7) and offers similar convergence rates as SAG. As part of the procedure, SDCA involves large primal gradient steps $\theta_{n-1} - (1/\mu)\nabla f^{\hat{t}_n}(\theta_{n-1})$ for updating the dual variables. It is thus appealing to study whether such large gradient steps can be used in (3.6) in the strongly convex case, regardless of the majorization-minimization principle. In other words, we want to study the use of the following surrogates within MISO:

$$g_n^t : \theta \mapsto f^t(\kappa_{n-1}^t) + \nabla f^t(\kappa_{n-1}^t)^\top(\theta - \kappa_{n-1}^t) + \frac{\mu}{2}\|\theta - \kappa_{n-1}^t\|_2^2, \tag{3.8}$$

which are lower bounds of the functions f^t instead of upper bounds. Then, minimizing $(1/T)\sum_{t=1}^T g_n^t$ amounts to performing the update (3.6) when replacing L by μ. The resulting algorithm is slightly different from SDCA but resembles it. As shown in the next proposition, the method achieves a fast convergence rate when $T \geq 2L/\mu$ but may diverge if T is small. Note that concurrently to our work, a similar result was independently obtained by Defazio, Caetano, and Domke [18], where a refined analysis provides a slightly better rate, namely, the constant 1/3 in (3.9) may be replaced by 1/2.

Proposition 3.4 (MISO for strongly convex unconstrained smooth problems). *Assume that the functions f^t are μ-strongly convex, L-smooth, and bounded below. Let $\theta^\star$ be a minimizer of f on Θ. Assume that $T \geq 2L/\mu$. When the functions g_n^t of* (3.8) *are used in Algorithm* 2*, we have for all $n \geq 1$,*

$$\mathbb{E}[f(\theta_n) - f^\star] \leq \left(1 - \frac{1}{3T}\right)^n \frac{2T}{\mu}\|\nabla f(\theta_0)\|_2^2. \tag{3.9}$$

When the functions f^t are lower-bounded by the function $\theta \mapsto (\mu/2)\|\theta\|_2^2$, we can use the initialization $\theta_0 = 0$ and $g_0^t : \theta \mapsto (\mu/2)\|\theta\|_2^2$ for all t. Then, the quantity $(2T/\mu)\|\nabla f(\theta_0)\|_2^2$ in (3.9) can be replaced by $Tf^\star$.

Proof. As in the proof of Proposition 3.1, we introduce the function $\bar{g}_n \triangleq \frac{1}{T}\sum_{t=1}^T g_n^t$, which is minimized by θ_n for $n \geq 1$. Since $\bar{g}_n$ is a lower bound on f, we have the relation $\bar{g}_n(\theta_n) \leq \bar{g}_n(\theta^\star) \leq f^\star$. Inspired by the convergence proof of SDCA [51], which computes an convergence rate of an expected duality gap, we proceed by studying the convergence of the sequence $(f^\star - \mathbb{E}[\bar{g}_n(\theta_n)])_{n\geq 1}$.

On the one hand, we have for all $n \geq 1$,

$$\begin{aligned}\bar{g}_n(\theta_n) &= \bar{g}_n(\theta_{n-1}) - \frac{\mu}{2}\|\theta_n - \theta_{n-1}\|_2^2 \\ &= \bar{g}_{n-1}(\theta_{n-1}) + \delta(g_n^{\hat{t}_n}(\theta_{n-1}) - g_{n-1}^{\hat{t}_n}(\theta_{n-1})) - \frac{\mu}{2}\|\theta_n - \theta_{n-1}\|_2^2,\end{aligned} \tag{3.10}$$

where $\delta = 1/T$. The first equality is true because $\bar{g}_n$ is quadratic and is minimized by θ_n, and the second one uses the relation (3.1). By definition of $g_n^{\hat{t}}$, we have that $g_n^{\hat{t}_n}(\theta_{n-1}) = f^{\hat{t}_n}(\theta_{n-1})$, and by taking the expectation, $\mathbb{E}[g_n^{\hat{t}_n}(\theta_{n-1})] = \mathbb{E}[f^{\hat{t}_n}(\theta_{n-1})] = \mathbb{E}[\mathbb{E}[f^{\hat{t}_n}(\theta_{n-1})|\mathcal{F}_{n-1}]] = \mathbb{E}[f(\theta_{n-1})]$, where $\mathcal{F}_n$ is the filtration representing all information up to iteration n. We also have that $\mathbb{E}[g_{n-1}^{\hat{t}_n}(\theta_{n-1})] = \mathbb{E}[\mathbb{E}[g_{n-1}^{\hat{t}_n}(\theta_{n-1})|\mathcal{F}_{n-1}]] = \mathbb{E}[\bar{g}_{n-1}(\theta_{n-1})]$. Thus, we obtain a first useful relation:

$$\mathbb{E}[\bar{g}_n(\theta_n)] = (1-\delta)\mathbb{E}[\bar{g}_{n-1}(\theta_{n-1})] + \delta\mathbb{E}[f(\theta_{n-1})] - \frac{\mu}{2}\mathbb{E}\left[\|\theta_n - \theta_{n-1}\|_2^2\right]. \tag{3.11}$$

On the other hand, for all $n \geq 2$,

$$\begin{aligned}\bar{g}_n(\theta_n) &= \bar{g}_{n-1}(\theta_n) + \delta(g_n^{\hat{t}_n}(\theta_n) - g_{n-1}^{\hat{t}_n}(\theta_n)) \\ &= \bar{g}_{n-1}(\theta_{n-1}) + \frac{\mu - \delta L}{2}\|\theta_n - \theta_{n-1}\|_2^2 + \delta\left(g_n^{\hat{t}_n}(\theta_n) + \frac{L}{2}\|\theta_n - \theta_{n-1}\|_2^2 - g_{n-1}^{\hat{t}_n}(\theta_n)\right) \\ &\geq \bar{g}_{n-1}(\theta_{n-1}) + \frac{\mu - \delta L}{2}\|\theta_n - \theta_{n-1}\|_2^2.\end{aligned} \tag{3.12}$$

We have used the fact that $\theta \mapsto g_n^{\hat{t}_n}(\theta) + (L/2)\|\theta - \theta_{n-1}\|_2^2$ is a majorizing surrogate of $f^{\hat{t}_n}$, whereas $g_{n-1}^{\hat{t}_n}$ is minorizing $f^{\hat{t}_n}$. By adding twice (3.12) after taking the expectation and once (3.11), we obtain that for all $n \geq 2$,

$$\begin{aligned}3\mathbb{E}[\bar{g}_n(\theta_n)] &\geq (3-\delta)\mathbb{E}[\bar{g}_{n-1}(\theta_{n-1})] + \delta\mathbb{E}[f(\theta_{n-1})] + \left(\frac{\mu}{2} - \delta L\right)\mathbb{E}[\|\theta_n - \theta_{n-1}\|_2^2] \\ &\geq (3-\delta)\mathbb{E}[\bar{g}_{n-1}(\theta_{n-1})] + \delta\mathbb{E}[f(\theta_{n-1})],\end{aligned} \tag{3.13}$$

where the second inequality comes from the large sample size condition $\delta L \leq \mu/2$. Since $\mathbb{E}[f(\theta_{n-1})] \geq f^\star$, this immediately gives for $n \geq 2$,

$$f^\star - \mathbb{E}\left[\bar{g}_n(\theta_n)\right] \leq \left(1 - \frac{1}{3T}\right)\left(f^\star - \mathbb{E}\left[\bar{g}_{n-1}(\theta_{n-1})\right]\right).$$

To obtain a convergence rate for $\mathbb{E}[f(\theta_n)] - f^\star$, we use again (3.13). For $n \geq 2$,

$$\begin{aligned}\delta(\mathbb{E}[f(\theta_{n-1})] - f^\star) &\leq \delta(\mathbb{E}[f(\theta_{n-1})] - \mathbb{E}[\bar{g}_{n-1}(\theta_{n-1})]) \\ &\leq 3(\mathbb{E}[\bar{g}_n(\theta_n)] - \mathbb{E}[\bar{g}_{n-1}(\theta_{n-1})]) \\ &\leq 3(f^\star - \mathbb{E}[\bar{g}_{n-1}(\theta_{n-1})]) \\ &\leq 3\left(1 - \frac{1}{3T}\right)^{n-2}(f^\star - \bar{g}_1(\theta_1)),\end{aligned} \tag{3.14}$$

and we obtain the convergence rate (3.9) by first noticing that

$$\begin{aligned}f^\star - \bar{g}_1(\theta_1) &= f^\star - \bar{g}_1(\theta_0) + \frac{\mu}{2}\|\theta_0 - \theta_1\|_2^2 \\ &= f^\star - f(\theta_0) + \frac{\mu}{2}\left\|\frac{1}{\mu}\nabla f(\theta_0)\right\|_2^2 \\ &\leq \frac{1}{2\mu}\|\nabla f(\theta_0)\|_2^2,\end{aligned}$$

where we use the relation $\bar{g}_1 = \bar{g}_0$ and $\bar{g}_0(\theta_0) = f(\theta_0)$. Then, we use the fact that $(1 - 1/3T) \geq 5/6$ since $T \geq 2L/\mu \geq 2$, such that $3(1-1/3T)^{-1}/(2\mu) \leq 9/(5\mu) \leq 2/\mu$.

To prove the last part of the proposition, we remark that all inequalities we have proved so far for $n \geq 2$ become true for $n = 1$. Thus, the last inequality in (3.14) is also true when replacing $n-2$ by $n-1$ and $\bar{g}_1(\theta_1)$ by $\bar{g}_0(\theta_0) = 0$. □

The proof technique is inspired in part by that for SDCA [51]; the quantity $\sum_{t=1}^T g_n^t(\theta_n)$ is indeed a lower bound of $f^\star$ and plays a similar role as the dual value in SDCA. We remark that the convergence rate (3.9) improves significantly upon the original one (3.3) and is similar to that of SAG when T is larger than $2L/\mu$.[3] However, Proposition 3.4 applies only to strongly convex problems. In other cases, the more conservative rule (3.6) should be preferred in theory, even though we present heuristics in section 3.4 that suggest using larger step sizes than $1/L$ in practice.

3.3. MISO for composite optimization. When f can be written as $f = (1/T)\sum_{t=1}^T f_1^t + f_2$, where the functions f_1^t are L-smooth, we can use the proximal gradient surrogate presented in section 2.3.2; it yields the following rule:

$$\theta_n \in \arg\min_{\theta\in\Theta} \frac{1}{2}\left\|\theta - \left(\frac{1}{T}\sum_{t=1}^T \kappa_{n-1}^t - \frac{1}{LT}\sum_{t=1}^T \nabla f_1^t(\kappa_{n-1}^t)\right)\right\|_2^2 + \frac{\lambda}{L} f_2(\theta), \tag{3.15}$$

where the vectors κ_{n-1}^t are defined as in section 3.2. This update is related to SDCA, as well as to stochastic methods for composite convex optimization such as the regularized dual averaging algorithm of Xiao [55]. As in the previous section, we obtain guarantees for nonconvex optimization, but our linear convergence rate for strongly convex problems is not as fast as that of SDCA. Even though we do not have a similar result as Proposition 3.4 for the composite setting, we have observed that using a smaller value for L than the theoretical one could work well in practice. We detail such an empirical strategy in the next section.

3.4. Practical implementation and heuristics. We have found the following strategies to improve the practical performance of MISO.

[3] Note that a similar assumption appears in the first analysis of SAG published in [36] before its refinement in [50].

Initialization. A first question is how to initialize the surrogates g_0^t in practice. Even though we have suggested the functions g_0^t to be in $\mathcal{S}_L(f^t, \theta_0)$ in Algorithm 2, our analysis weakly relies on this assumption. In fact, most of our results hold when choosing surrogates computed at points κ_0^t that are not necessarily equal to θ_0; at most only constants from the convergence rates would be affected by such a change. An effective empirical strategy is inspired by the second part of Proposition 3.4: we first define functions $g_0^t : \theta \mapsto (L/2)\|\theta - \theta_0\|_2^2$ and perform T iterations of MISO without randomization, selecting the function f^t at iteration t, such that each surrogate is updated exactly once. Then, we use these updated surrogates for initializing the regular randomized scheme.

Warm restart and continuation. When available, warm restart can be used for initializing the surrogates. Assume that we are interested in minimizing a composite function $(1/T)\sum_{t=1}^T f_1^t(\theta) + \lambda f_2(\theta)$, which is parameterized by a scalar λ, and that we want to obtain a minimizer for several parameter values $\lambda_1 < \lambda_2 < \cdots < \lambda_M$. We first solve the problem for $\lambda = \lambda_M$ and then use the surrogates obtained at the end of the optimization for initializing the algorithm when addressing the problem with $\lambda = \lambda_{M-1}$. We proceed similarly going from larger to smaller values of λ. We have empirically observed that the warm restart strategy could be extremely efficient in practice and would deserve further study in a future work.

Heuristics for selecting step sizes. Choosing proximal gradient surrogates g^t requires choosing some Lipschitz constant L (or a strong convexity parameter μ for Proposition 3.4), which leads to a specific step size in (3.15). However, finding an appropriate step size can be difficult in practice for several reasons: (i) in some cases, these parameters are unknown; (ii) even though a global Lipschitz constant might be available, a local Lipschitz constant could be more effective; (iii) the convergence rates of Proposition 3.3 can be obtained by choosing a smaller value for L than the "true" Lipschitz constant, as long as the inequality $\mathbb{E}[f(\theta_n)] \leq \mathbb{E}[\bar{g}_n(\theta_n)]$ is always satisfied, where $\bar{g}_n \triangleq (1/T)\sum_{t=1}^T g_n^t$. This motivates the following heuristics:

MISO1 First perform one pass over $\eta = 5\%$ of the data to select a constant $L_1 = 2^{-k}L_0$ with k chosen among positive integers, yielding the smallest objective on the data subset, where L_0 is an upper bound of the true Lipschitz constant.

MISO2 Proceed as in MISO1, but choose a more aggressive strategy $L_2 = L_1\eta$; during the optimization, compute the quantities a_n^t and b_n^t defined as $a_n^t = a_{n-1}^t$, $b_n^t = b_{n-1}^t$ if $t \neq \hat{t}_n$, and otherwise $a_n^{\hat{t}_n} = f^{\hat{t}_n}(\theta_{n-1})$, $b_n^{\hat{t}_n} = g_{L_2}^{\hat{t}_n}(\theta_{n-1})$, where we have parameterized the surrogates g^t by L_2. Every T iterations, compare the sums $A_n = \sum_{t=1}^T a_n^t$ and $B_n = \sum_{t=1}^T b_n^t$. If $A_n \leq B_n$, do nothing; otherwise, increase the value of L_2 until this inequality is satisfied.

The heuristic MISO2 is more aggressive than MISO1 since it starts with a smaller value for L. After every iteration, this value is possibly increased such that on average, the surrogates "behave" as majorizing functions. Even though this heuristic does not come with any theoretical guarantee, it was found to perform slightly better than MISO1 for strongly convex problems.

Using a different parameter L_t for every function f_t. Even though our analysis was conducted with a global parameter L for simplicity, it is easy to extend the analysis when the parameter L is adjusted individually for every surrogate. This is useful when the functions f_t are heterogeneous.

Parallelization with mini-batches. The complexity of MISO is often dominated by the cost of updating the surrogates $g_n^{\hat{t}_n}$, which typically requires computing the gradient of a function. A simple extension is to update several surrogates at the same time when parallel computing facilities are available.

TABLE 1
Description of datasets used in our experiments.

Name	T	p	Storage	Density	Size (GB)
covtype	581 012	54	dense	1	0.23
alpha	500 000	500	dense	1	1.86
ocr	2 500 000	1 155	dense	1	21.5
real-sim	72 309	20 958	sparse	0.0024	0.056
rcv1	781 265	47 152	sparse	0.0016	0.89
webspam	250 000	16 091 143	sparse	0.0002	13.90

4. Experimental validation. In this section, we evaluate MISO on large-scale machine learning problems. Our implementation is coded in C++ interfaced with MATLAB and is freely available in the open-source software package SPAMS [40].[4] All experiments were conducted on a single core of a 2-GHz Intel CPU with 64 GB of RAM.

Datasets. We use six publicly available datasets, which consist of pairs $(y_t, \mathbf{x}_t)_{t=1}^T$, where the y_t's are labels in $\{-1, +1\}$ and the $\mathbf{x}_t$'s are vectors in $\mathbb{R}^p$ representing data points. The datasets are described in Table 1. alpha, rcv1, ocr, and webspam are obtained from the 2008 Pascal large-scale learning challenge.[5] covtype and real-sim are obtained from the LIBSVM website.[6] The datasets are preprocessed as follows: all dense datasets are standardized to have zero-mean and unit variance for every feature. The sparse datasets are normalized such that each $\mathbf{x}_t$ has unit ℓ_2-norm.

4.1. ℓ_2-logistic regression. We consider the ℓ_2-regularized logistic regression problem, which can be formulated as follows:

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{T} \sum_{t=1}^{T} \ell(y_t, \mathbf{x}_t^\top \theta) + \frac{\lambda}{2} \|\theta\|_2^2, \tag{4.1}$$

where $\ell(u, \hat{u}) = \log(1 + e^{-u\hat{u}})$ for all $(u, \hat{u})$. Following [50], we report some results obtained with different methods with the parameter $\lambda = 1/T$, which is argued to be of the same order of magnitude as the smallest value that would be used in practice for machine learning problems. We also performed experiments with the values $\lambda = 0.1/T$ and $\lambda = 10/T$ to study the impact of the strong convexity parameter; the output of these two additional experiments is not reported in the present paper for space limitation reasons, but it will be discussed and taken into account in our conclusions. The algorithms included in the comparison are the following:

SGD-h The stochastic gradient descent algorithm with a heuristic for choosing the step size similar to MISO1 and inspired by Leon Bottou's sgd toolbox for machine learning.[7] A step size of the form $\rho/\sqrt{n + n_0}$ is automatically adjusted when performing one pass on $\eta = 5\%$ of the training data. We obtain consistent results with the performance of SGD reported by Schmidt, Le Roux, and Bach [50] when the step size is chosen from hindsight. Based on their findings, we do not include in our figures other variants of SGD, e.g., [21, 27, 29, 55].

FISTA The accelerated gradient method proposed by Beck and Teboulle [3] with a line-search for automatically adjusting the Lipschitz constant.

[4] http://spams-devel.gforge.inria.fr/.
[5] http://largescale.ml.tu-berlin.de.
[6] http://www.csie.ntu.edu.tw/~cjlin/libsvm/.
[7] http://leon.bottou.org/projects/sgd.

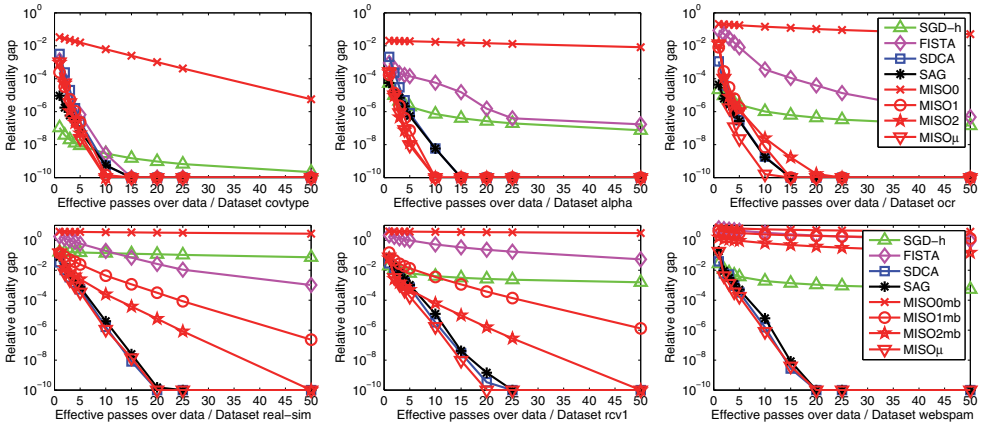

FIG. 2. *Relative duality gap obtained for logistic regression with respect to the number of passes over the data.*

SDCA The algorithm of Shalev-Schwartz and Zhang [51], efficiently implemented in the language C by Mark Schmidt.[8]

SAG A fast implementation in C also provided by Schmidt, Le Roux, and Bach [50]. We use the step size $1/L$ since it performed similarly to their heuristic line search.

MISO0 The majorization-minimization algorithm MISO, using the trivial upper bound $L^t = 0.25\|\mathbf{x}_t\|_2^2$ on the Lipschitz constant for example t.

MISO1 The majorization-minimization heuristic MISO1 described in section 3.4.

MISO2 The heuristic MISO2, also described in section 3.4.

MISOμ The update rule corresponding to Proposition 3.4.

For sparse datasets, MISO0, MISO1, and MISO2 are not practical since they suffer from a $O(Tp)$ memory cost. Their update rules can indeed be rewritten

$$\theta_n \leftarrow \theta_{n-1} - \frac{1}{T}\left(\left(\theta_{n-1} - \frac{1}{L}\nabla f^{\hat{t}_n}(\theta_{n-1})\right) - \left(\kappa_{n-1}^{\hat{t}_n} - \frac{1}{L}\nabla f^{\hat{t}_n}(\kappa_{n-1}^{\hat{t}_n})\right)\right),$$

where $f^t : \theta \mapsto \ell(y_t, \mathbf{x}_t^\top \theta) + \frac{\lambda}{2}\|\theta\|_2^2$. Thus, for every example t, the algorithm requires storing the dense vector $\kappa_{n-1}^t - (1/L)\nabla f^t(\kappa_{n-1}^t)$. Therefore, we use mini-batches of size $\lfloor 1/d \rfloor$, where d is the density of the dataset; the resulting algorithms, which we denote by MISO0-mb, MISO1-mb, and MISO2-mb, have a storage cost equal to $O(dpT)$, which is the same as the dataset.

On the other hand, the update rule MISOμ applied to the λ-strongly convex functions f^t admits a simpler and computationally cheaper form. Since $\kappa_{n-1}^t - (1/\lambda)\nabla f^t(\kappa_{n-1}^t) = -(1/\lambda)\ell'(y_t, \mathbf{x}_t^\top \kappa_{n-1}^t)\mathbf{x}_t$, the update becomes

$$\theta_n \leftarrow \theta_{n-1} - \frac{1}{T\lambda}\left(\ell'(y_{\hat{t}_n}, \mathbf{x}_{\hat{t}_n}^\top \theta_{n-1}) - \ell'(y_{\hat{t}_n}, \mathbf{x}_{\hat{t}_n}^\top \kappa_{n-1}^{\hat{t}_n})\right)\mathbf{x}_{\hat{t}_n}, \tag{4.2}$$

where ℓ' denotes the derivative of ℓ with respect to its second argument. Assuming that the dataset fits into memory, the only extra quantities to store are the scalars $\ell'(y_{\hat{t}_n}, \mathbf{x}_{\hat{t}_n}^\top \kappa_{n-1}^{\hat{t}_n})$, and the resulting memory cost is simply $O(T)$.

[8] http://www.di.ens.fr/~mschmidt/Software/SAG.html.

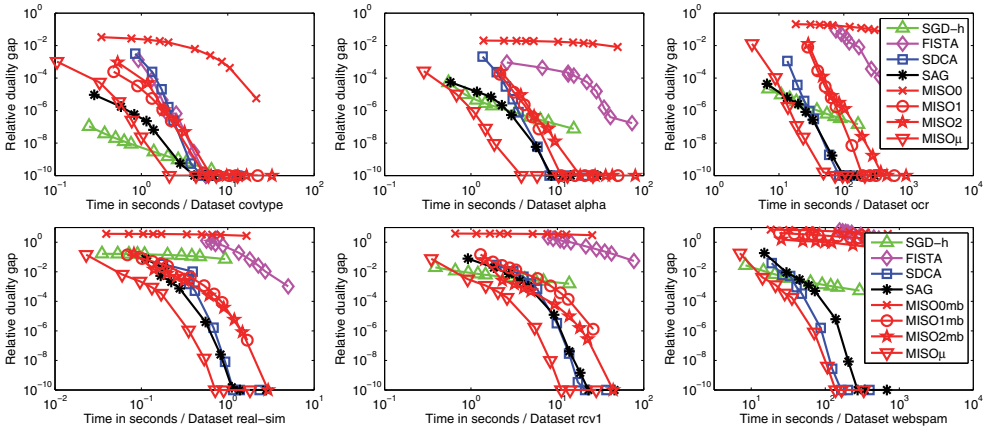

FIG. 3. *Relative duality gap obtained for logistic regression with respect to the CPU time.*

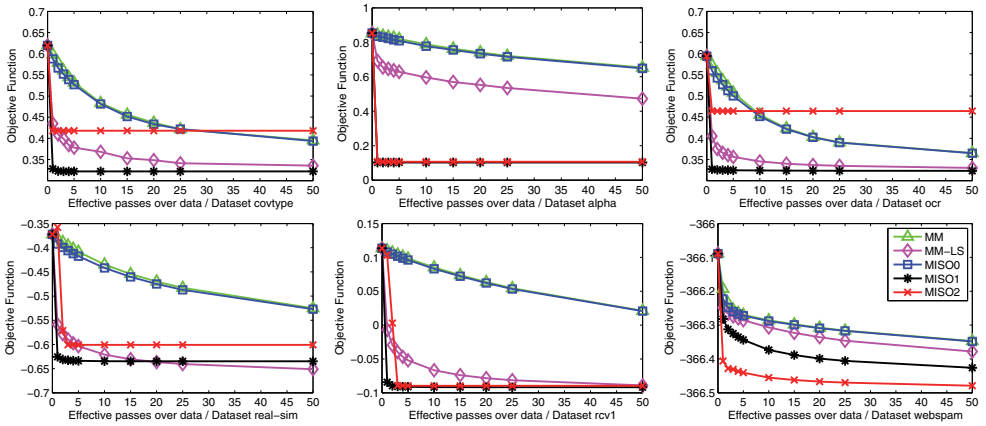

FIG. 4. *Objective function during the sparse estimation experiment.*

We present our comparison of the above methods with $\lambda = 1/T$ in Figures 2 and 3, where we plot the relative duality gap defined as $(f(\theta_n) - g^\star)/g^\star$, where $g^\star$ is the best value of the Fenchel dual that we have obtained during our experiments. The conclusions of our empirical study are the following:

- *SAG, SDCA, and MISOμ.* These methods perform similarly and were consistently the fastest, except in the regime $T < 2L/\mu$ where MISOμ can diverge.
- *The four variants of MISO.* As predicted by its theoretical convergence rate, MISO0 does not perform better than ISTA [3] without line search (not reported in the figures). MISO1 and MISO2 perform significantly better. MISOμ is always better than or as good as MISO1 and MISO2, except for sparse datasets with $\lambda = 0.1/T$ where the condition $T \geq 2L/\mu$ is not satisfied.
- *Influence of mini-batch.* Whereas MISO2 performs equally well as SAG/SDCA for dense datasets, mini-batches for sparse datasets makes it slower.
- *Stochastic gradient descent.* SGD-h always performs well at the beginning of the procedure but is not competitive compared to incremental approaches after a few passes over the data.

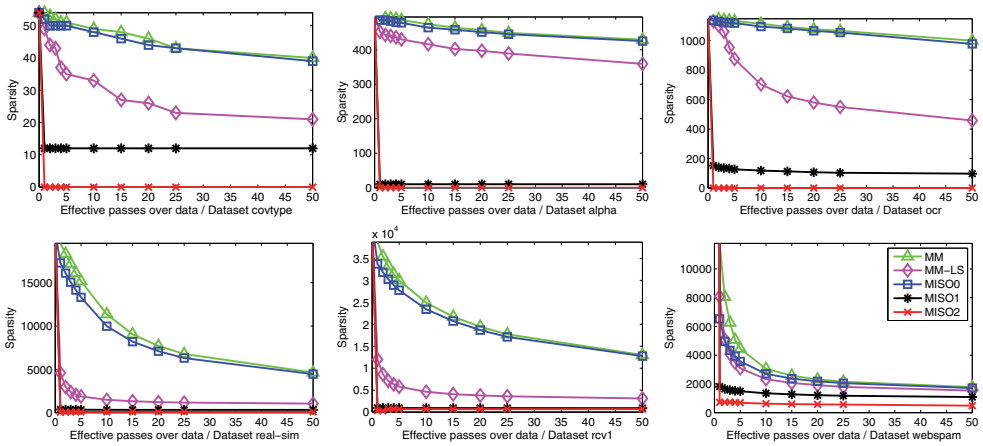

FIG. 5. *Sparsity of the solution during the sparse estimation experiment.*

Note that an evaluation of a preliminary version of MISO2 is presented in [38] for the ℓ_1-regularized logistic regression problem, where the objective function is not strongly convex. Our experimental findings showed that MISO2 was competitive with state-of-the-art solvers based on active-set and coordinate descent algorithms [24].

4.2. Nonconvex sparse estimation. The majorization-minimization principle is appealing for nonconvex and nonsmooth optimization, where only a few algorithms apply. Here, we address a sparse estimation problem presented in section 2.3.3:

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{T} \sum_{t=1}^{T} \frac{1}{2}(y_t - \mathbf{x}_t^\top \theta)^2 + \lambda \sum_{j=1}^{p} \log(|\theta[j]| + \varepsilon), \tag{4.3}$$

where the scalars y_t and the vectors $\mathbf{x}_t$ are the same as in the previous section, and ε is set to 0.01. The model parameter λ controls the sparsity of the solution. Even though (4.3) is nonconvex and nonsmooth, stationary points can be obtained in various ways. In this section, we consider majorization-minimization approaches where the penalty function $\theta \mapsto \sum_{j=1}^{p} \log(|\theta[j]| + \varepsilon)$ is upper-bounded as in (2.7), whereas the functions $\theta \mapsto (1/2)(y_t - \mathbf{x}_t^\top \theta)^2$ are upper-bounded by the Lipschitz gradient surrogates of section 2.3.1. We compare five approaches:

MM Algorithm 1 with the trivial Lipschitz constant $L = (1/T) \sum_{t=1}^{T} 0.25 \|\mathbf{x}_t\|_2^2$.

MM-LS Algorithm 1 with the line-search scheme of ISTA [3] for adjusting L.

MISO We compare MISO0, MISO1, and MISO2, as in the previous section.

We choose a parameter λ for each dataset, such that the solution with the lowest objective function obtained by any of the tested methods has approximately a sparsity of 10 for datasets covtype and alpha, 100 for ocr and real-sim, and 1 000 for rcv1 and webspam. The methods are initialized with $\theta_0 = (\|\mathbf{y}\|_2 / \|\mathbf{X}\mathbf{X}^\top \mathbf{y}\|_2)\mathbf{X}^\top \mathbf{y}$; indeed, the initialization $\theta_0 = 0$ that was a natural choice in section 4.1 appears to be often a bad stationary point of problem (4.3) and thus an inappropriate initial point. We report the objective function values for different passes over the data in Figure 4 and the sparsity of the solution in Figure 5. Our conclusions are the following:

- Methods with line searches do significantly better than those without, showing that adjusting the constant L is important for these datasets.
- MISO1 does asymptotically better than MM-LS for five of the datasets after 50 epochs and slightly worse for real-sim; in general, MISO1 seems to converge

substantially faster than other approaches, both in terms of objective function and in terms of the support of the solution.
- The performance of MISO2 is mitigated. In one case, it does better than MISO1, but in some others, it converges to the stationary point $\theta = 0$.

5. Conclusion. In this paper, we have presented new algorithms based on the majorization-minimization principle for minimizing a large sum of functions. The main asset of our approach is probably its applicability to a large class of nonconvex problems, including nonsmooth ones, where we obtain convergence and asymptotic stationary point guarantees. For convex problems, we also propose new incremental rules for composite optimization, which are competitive with state-of-the-art solvers in the context of large-scale machine learning problems such as logistic regression.

We note that other majorization-minimization algorithms have recently been analyzed, such as block coordinate variants in [38, 47] and stochastic ones in [12, 39, 48]. In particular, we have proposed in [39] a stochastic majorization-minimization algorithm that does not require storing information about past iterates, when the objective function is an expectation. Since the first version of our work was published in [39], MISO has also been extended by other authors in [57] using the alternating direction method of multipliers framework.

For future work, we are currently investigating extensions of the scheme MISOμ for strongly convex objective functions. We believe that the algorithm can be modified to remove the large sample condition $T \geq 2L/\mu$, that the convergence proof can be extended to the proximal setting, and that it is possible to use acceleration techniques in the sense of Nesterov [44]. Another interesting direction of research would be to study the stability of our result to inexact minimization of surrogate functions following for instance the analysis of [49] for proximal gradient methods.

Appendix A. Basic definitions and useful results. The following definitions can be found in classical textbooks, e.g, [5, 8, 10, 46]. For the sake of completeness, we briefly introduce them here.

DEFINITION A.1 (directional derivative). *Let us consider a function $f : \mathbb{R}^p \to \mathbb{R}$ and θ, θ' be in $\mathbb{R}^p$. When it exists, the following limit is called the directional derivative of f at θ in the direction $\theta' - \theta$: $\nabla f(\theta, \theta' - \theta) \triangleq \lim_{t \to 0^+} (f(\theta + t(\theta' - \theta)) - f(\theta))/t$. When f is differentiable at θ, directional derivatives exist in every direction, and $\nabla f(\theta, \theta' - \theta) = \nabla f(\theta)^\top (\theta' - \theta)$.*

DEFINITION A.2 (stationary point). *Let us consider a function $f : \Theta \subseteq \mathbb{R}^p \to \mathbb{R}$, where Θ is a convex set, such that f admits a directional derivative $\nabla f(\theta, \theta' - \theta)$ for all θ, θ' in Θ. We say that θ in Θ is a stationary point if for all θ' in Θ, $\nabla f(\theta, \theta' - \theta) \geq 0$.*

DEFINITION A.3 (Lipschitz continuity). *A function $f : \mathbb{R}^p \to \mathbb{R}$ is L-Lipschitz continuous for some $L > 0$ when for all θ, θ' in $\mathbb{R}^p$, $|f(\theta') - f(\theta)| \leq L \|\theta - \theta'\|_2$.*

DEFINITION A.4 (strong convexity). *Let Θ be a convex set. A function $f : \Theta \subseteq \mathbb{R}^p \to \mathbb{R}$ is called μ-strongly convex when there exists a constant $\mu > 0$ such that for all θ' in Θ, the function $\theta \mapsto f(\theta) - \frac{\mu}{2}\|\theta - \theta'\|_2^2$ is convex.*

We now present two lemmas that are useful for characterizing first-order surrogate functions. Their proofs can be found in the appendix of [38].

LEMMA A.5 (regularity of residual functions). *Let $f, g : \mathbb{R}^p \to \mathbb{R}$ be two functions. Define $h \triangleq g - f$. Then, if g is ρ-strongly convex and f is L-smooth, with $\rho \geq L$, h is $(\rho - L)$-strongly convex; if g and f are convex and L-smooth, h is also L-smooth; if g and f are μ-strongly convex and L-smooth, h is $(L - \mu)$-smooth.*

LEMMA A.6 (Regularity of optimal value functions). *Let $f : \mathbb{R}^{p_1} \times \Theta_2 \to \mathbb{R}$ be a function of two variables where $\Theta_2 \subseteq \mathbb{R}^{p_2}$ is a convex set. Assume that*

- $\theta_1 \mapsto f(\theta_1, \theta_2)$ *is differentiable for all* θ_2 *in* Θ_2*;*
- $\theta_2 \mapsto \nabla_1 f(\theta_1, \theta_2)$ *is* L*-Lipschitz continuous for all* θ_1 *in* $\mathbb{R}^{p_1}$*;*
- $\theta_2 \mapsto f(\theta_1, \theta_2)$ *is* μ*-strongly convex for all* θ_1 *in* $\mathbb{R}^{p_1}$*.*

Also define $\tilde{f}(\theta_1) \triangleq \min_{\theta_2 \in \Theta_2} f(\theta_1, \theta_2)$*. Then,* $\tilde{f}$ *is differentiable and* $\nabla \tilde{f}(\theta_1) = \nabla_1 f(\theta_1, \theta_2^\star)$*, where* $\theta_2^\star \triangleq \arg\min_{\theta_2 \in \Theta_2} f(\theta_1, \theta_2)$*. Moreover, if* $\theta_1 \mapsto \nabla_1 f(\theta_1, \theta_2)$ *is* L'*-Lipschitz continuous for all* θ_1 *in* $\mathbb{R}^{p_1}$*, the gradient* $\nabla \tilde{f}$ *is* $(L' + L^2/\mu)$*-Lipschitz.*

Acknowledgments. The author would like to thank Zaid Harchaoui, Francis Bach, Simon Lacoste-Julien, Mark Schmidt, Martin Jaggi, the associate editor, and the anonymous reviewers for their useful comments.

REFERENCES

[1] S. Ahn, J. A. Fessler, D. Blatt, and A. O. Hero, *Convergent incremental optimization transfer algorithms: Application to tomography*, IEEE Trans. Med. Imaging, 25 (2006), pp. 283–296.

[2] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, *Optimization with sparsity-inducing penalties*, Found. Trends Mach. Learn., 4 (2012), pp. 1–106.

[3] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.

[4] A. Beck and L. Tetruashvili, *On the convergence of block coordinate descent type methods*, SIAM J. Optim., 23 (2013), pp. 2037–2060.

[5] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed., Athena Scientific, Belmont, MA, 1999.

[6] D. Blatt, A. O. Hero, and H. Gauchman, *A convergent incremental gradient method with a constant step size*, SIAM J. Optim., 18 (2007), pp. 29–51.

[7] D. Böhning and B. G. Lindsay, *Monotonicity of quadratic-approximation algorithms*, Ann. Inst. Statist. Math., 40 (1988), pp. 641–663.

[8] J. M. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization: Theory and Examples*, Springer, New York, 2006.

[9] L. Bottou, *Online algorithms and stochastic approximations*, in Online Learning and Neural Networks, D. Saad, ed., Cambridge University Press, Cambridge, UK, 1998.

[10] S. P. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[11] E. J. Candès, M. Wakin, and S. P. Boyd, *Enhancing sparsity by reweighted ℓ_1 minimization*, J. Fourier Anal. Appl., 14 (2008), pp. 877–905.

[12] A. Choromanska and T. Jebara, *Stochastic Bound Majorization*, arXiv:1309.5605, 2013.

[13] M. Collins, R. E. Schapire, and Y. Singer, *Logistic regression, AdaBoost and Bregman distances*, Mach. Learn., 48 (2002), pp. 253–285.

[14] P. L. Combettes and J.-C. Pesquet, *Proximal splitting methods in signal processing*, in Fixed-Point Algorithms for Inverse Problems in Science and Engineering, Springer, New York, 2010.

[15] P. L. Combettes and V. R. Wajs, *Signal recovery by proximal forward-backward splitting*, Multiscale Model. Simul., 4 (2005), pp. 1168–1200.

[16] I. Daubechies, M. Defrise, and C. De Mol, *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Comm. Pure Appl. Math., 57 (2004), pp. 1413–1457.

[17] A. J. Defazio, F. Bach, and S. Lacoste-Julien, *SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives*, in Proceedings of Advances in Neural Information Processing Systems, 2014.

[18] A. J. Defazio, T. S. Caetano, and J. Domke, *Finito: A faster, permutable incremental gradient method for big data problems*, in Proceedings of ICML, 2014.

[19] S. Della Pietra, V. Della Pietra, and J. Lafferty, *Duality and Auxiliary Functions for Bregman Distances*, Tech. report, CMU-CS-01-109, Carnegie Mellon University, Pittsburgh, 2001.

[20] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. Roy. Statist. Soc. Ser. B, 39 (1977), pp. 1–38.

[21] J. Duchi, E. Hazan, and Y. Singer, *Adaptive subgradient methods for online learning and stochastic optimization*, J. Mach. Learn. Res., 12 (2011), pp. 2121–2159.

[22] J. Duchi and Y. Singer, *Efficient online and batch learning using forward backward splitting*, J. Mach. Learn. Res., 10 (2009), pp. 2899–2934.
[23] H. Erdogan and J. A. Fessler, *Ordered subsets algorithms for transmission tomography*, Phys. Med. Biol., 44 (1999), pp. 2835–2851.
[24] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, *LIBLINEAR: A library for large linear classification*, J. Mach. Learn. Res., 9 (2008), pp. 1871–1874.
[25] M. Fashing and C. Tomasi, *Mean shift is a bound optimization*, IEEE Trans. Pattern Anal., 27 (2005), pp. 471–474.
[26] G. Gasso, A. Rakotomamonjy, and S. Canu, *Recovering sparse signals with non-convex penalties and DC programming*, IEEE Trans. Signal Process., 57 (2009), pp. 4686–4698.
[27] S. Ghadimi and G. Lan, *Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization* I*: A generic algorithmic framework*, SIAM J. Optim., 22 (2012), pp. 1469–1492.
[28] E. T. Hale, W. Yin, and Y. Zhang, *Fixed-point continuation for ℓ_1-minimization: Methodology and convergence*, SIAM J. Optim., 19 (2008), pp. 1107–1130.
[29] E. Hazan and S. Kale, *Beyond the regret minimization barrier: An optimal algorithm for stochastic strongly-convex optimization*, in Proceedings of COLT, 2011.
[30] R. Horst and N. V. Thoai, *DC programming: Overview*, J. Optim. Theory Appl., 103 (1999), pp. 1–43.
[31] T. Jebara and A. Choromanska, *Majorization for CRFs and latent likelihoods*, in Proceedings of Advances in Neural Information Processing Systems, 2012.
[32] A. Juditsky and A. Nemirovski, *First order methods for nonsmooth convex large-scale optimization*, in Optimization for Machine Learning, MIT Press, Cambridge, MA, 2011.
[33] E. Khan, B. Marlin, G. Bouchard, and K. Murphy, *Variational bounds for mixed-data factor analysis*, in Proceedings of Advances in Neural Information Processing Systems, 2010.
[34] G. Lan, *An optimal method for stochastic composite optimization*, Math. Program., 133 (2012), pp. 365–397.
[35] K. Lange, D. R. Hunter, and I. Yang, *Optimization transfer using surrogate objective functions*, J. Comput. Graph. Statist., 9 (2000), pp. 1–20.
[36] N. Le Roux, M. Schmidt, and F. Bach, *A stochastic gradient method with an exponential convergence rate for finite training sets*, in Proceedings of Advances in Neural Information Processing Systems, 2012.
[37] D. D. Lee and H. S. Seung, *Algorithms for non-negative matrix factorization*, in Proceedings of Advances in Neural Information Processing Systems, 2001.
[38] J. Mairal, *Optimization with first-order surrogate functions*, in Proceedings of ICML, 2013.
[39] J. Mairal, *Stochastic majorization-minimization algorithms for large-scale optimization*, in Proceedings of Advances in Neural Information Processing Systems, 2013.
[40] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, *Online learning for matrix factorization and sparse coding*, J. Mach. Learn. Res., 11 (2010), pp. 19–60.
[41] J. J. Moreau, *Fonctions convexes duales et points proximaux dans un espace hilbertien*, C. R. Acad. Sci. Paris Sér. A Math., 255 (1962), pp. 2897–2899.
[42] R. M. Neal and G. E. Hinton, *A view of the EM algorithm that justifies incremental, sparse, and other variants*, in Learning in Graphical Models, Kluwer, Dordrecht, the Netherlands, 1998, pp. 355–368.
[43] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optim., 19 (2009), pp. 1574–1609.
[44] Y. Nesterov, *Introductory Lectures on Convex Optimization*, Kluwer, Dordrecht, the Netherlands, 2004.
[45] Y. Nesterov, *Gradient methods for minimizing composite objective functions*, Math. Program., 140 (2012), pp. 125–161.
[46] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed., Springer, New York, 2006.
[47] M. Razaviyayn, M. Hong, and Z.-Q. Luo, *A unified convergence analysis of block successive minimization methods for nonsmooth optimization*, SIAM J. Optim., 23 (2013), pp. 1126–1153.
[48] M. Razaviyayn, M. Sanjabi, and Z.-Q. Luo, *A Stochastic Successive Minimization Method for Nonsmooth Nonconvex Optimization*, arXiv:1307.4457v2, 2013.
[49] M. Schmidt, N. Le Roux, and F. Bach, *Convergence rates of inexact proximal-gradient methods for convex optimization*, in Proceedings of Advances in Neural Information Processing Systems, 2011.
[50] M. Schmidt, N. Le Roux, and F. Bach, *Minimizing Finite Sums with the Stochastic Average Gradient*, arXiv:1309.2388, 2013.

[51] S. Shalev-Schwartz and T. Zhang, *Proximal Stochastic Dual Coordinate Ascent*, arXiv:1211.2717, 2012.
[52] B. A. Turlach, W. N. Venables, and S. J. Wright, *Simultaneous variable selection*, Technometrics, 47 (2005), pp. 349–363.
[53] M. J. Wainwright and M. I. Jordan, *Graphical models, exponential families, and variational inference*, Found. Trends Mach. Learn., 1 (2008), pp. 1–305.
[54] S. J. Wright, R. D. Nowak, and M. A. Figueiredo, *Sparse reconstruction by separable approximation*, IEEE Trans. Signal Process., 57 (2009), pp. 2479–2493.
[55] L. Xiao, *Dual averaging methods for regularized stochastic learning and online optimization*, J. Mach. Learn. Res., 11 (2010), pp. 2543–2596.
[56] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables.*, J. Roy. Statist. Soc. Ser. B, 68 (2006), pp. 49–67.
[57] L. W. Zhong and J. T. Kwok, *Fast stochastic alternating direction method of multipliers*, in Proceedings of ICML, 2014.